

Comparison of complete genomes: Functional and evolutionary inferences

Eugene V. Koonin

National Center for Biotechnology Information, NLM, NIH

- Contents:**

- The status of complete genome sequencing**
- Overview of computer methods for genome sequence analysis**
- The genome of an intracellular parasitic bacterium - *Chlamydia trachomatis*- a case study**
- Phylogenetic classification of proteins from complete genomes**
- Structural genomics**
- Some important tools for genome analysis**

Microbiology in the 1980s - gene-centered approach

- Perform a mutant screen
- Obtain mutants of interest
- Do genetics to map the gene
- Clone the gene
- Study the gene product and make inferences

What happened after 1995?

- All of a sudden ~20 microbial genomes have been sequenced and we have many more genes to study than was ever previously imaginable. Better yet, these genes are organized in complete, genome-specific sets.

How much data does the public have?

- B *Haemophilus influenzae*
- B *Mycoplasma genitalium*
- B *Synechocystis* sp.
- B *Mycoplasma pneumoniae*
- B *Helicobacter pylori* (gastric ulcers)
- B *Escherichia coli*
- B *Bacillus subtilis*
- B *Borrelia burgdorferi* (Lyme's Disease)
- B *Aquifex aeolicus*
- B *Mycobacterium tuberculosis* (TB)
- B *Treponema pallidum* (Syphilis)
- B *Chlamydia trachomatis* (STD)
- B *Rickettsia prowazekii* (typhus)

- A *Methanococcus jannaschii*
- A *Methanobacterium thermoautotrophicum*
- A *Archaeoglobus fulgidus*
- A *Pyrococcus horikoshii*

- E *Saccharomyces cerevisiae*
- E *Caenorhabditis elegans*
- E *Plasmodium falciparum* (Malaria)- 2 chromosomes

A- Archaea

B- Bacteria

E- Eukarya

The pathogenic
microbes are in red.

Information on complete genomes is collected at the NCBI

The screenshot shows a Netscape browser window titled "Entrez genomes - Netscape". The address bar displays the URL "http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html". The browser's menu bar includes "File", "Edit", "View", "Go", "Communicator", and "Help". Below the menu bar, there are icons for "Bookmarks", "Instant Message", "Internet", "Lookup", and "New&Cool".

The main content area features the NCBI logo and the title "Entrez Genomes". A navigation bar contains links to "NCBI", "BLAST", "Nucleotides", "Proteins", "Structure", "Taxonomy", "PubMed", and "Help". Below this is a search bar with the label "Search_for:" and a dropdown menu set to "All Fields".

A sidebar on the left lists various categories: "All Organisms", "Prominent Organisms", "Microbial genomes" (with sub-links for "BLAST" and "List of projects"), "Archaea" (with sub-links for "Genome" and "Plasmids"), "Bacteria" (with sub-links for "Genome" and "Plasmids"), and "Eukaryota" (with sub-links for "Genome", "Plasmids", and "Organelles").

The main content area is titled "Prominent Organisms Taxonomy / List". It lists several complete genomes under the heading "Complete Genomes":

- Aquifex aeolicus*
- Archaeoglobus fulgidus*
- Bacillus subtilis*
- Borrelia burgdorferi*
 - chromosome
 - plasmids: *cp26*, *cp9*, *lp17*, *lp25*, *lp28-1*, *lp28-2*, *lp28-3*, *lp28-4*, *lp36*, *lp38*, *lp54*
- Chlamydia trachomatis*
- Chlamydia pneumoniae* ^{NEW}
- Escherichia coli*
- Haemophilus influenzae*
- Helicobacter pylori*
- Helicobacter pylori J99*
- Methanobacterium thermoautotrophicum*
- [3] *Methanococcus jannaschii*
 - chromosome
 - small extrachromosomal element
 - large extrachromosomal element
- Mycobacterium tuberculosis*
- Mycoplasma genitalium*
- Mycoplasma pneumoniae*
- Pyrococcus horikoshii*
- Rickettsia prowazekii*
- [16] *Saccharomyces cerevisiae*
 - chromosomes: I, II, III, IV, V, VI, VII, VIII, VIII, IX, X, XI, XII, XIII, XIV,

How do we deal with this data explosion-

1) Experimental Approaches-

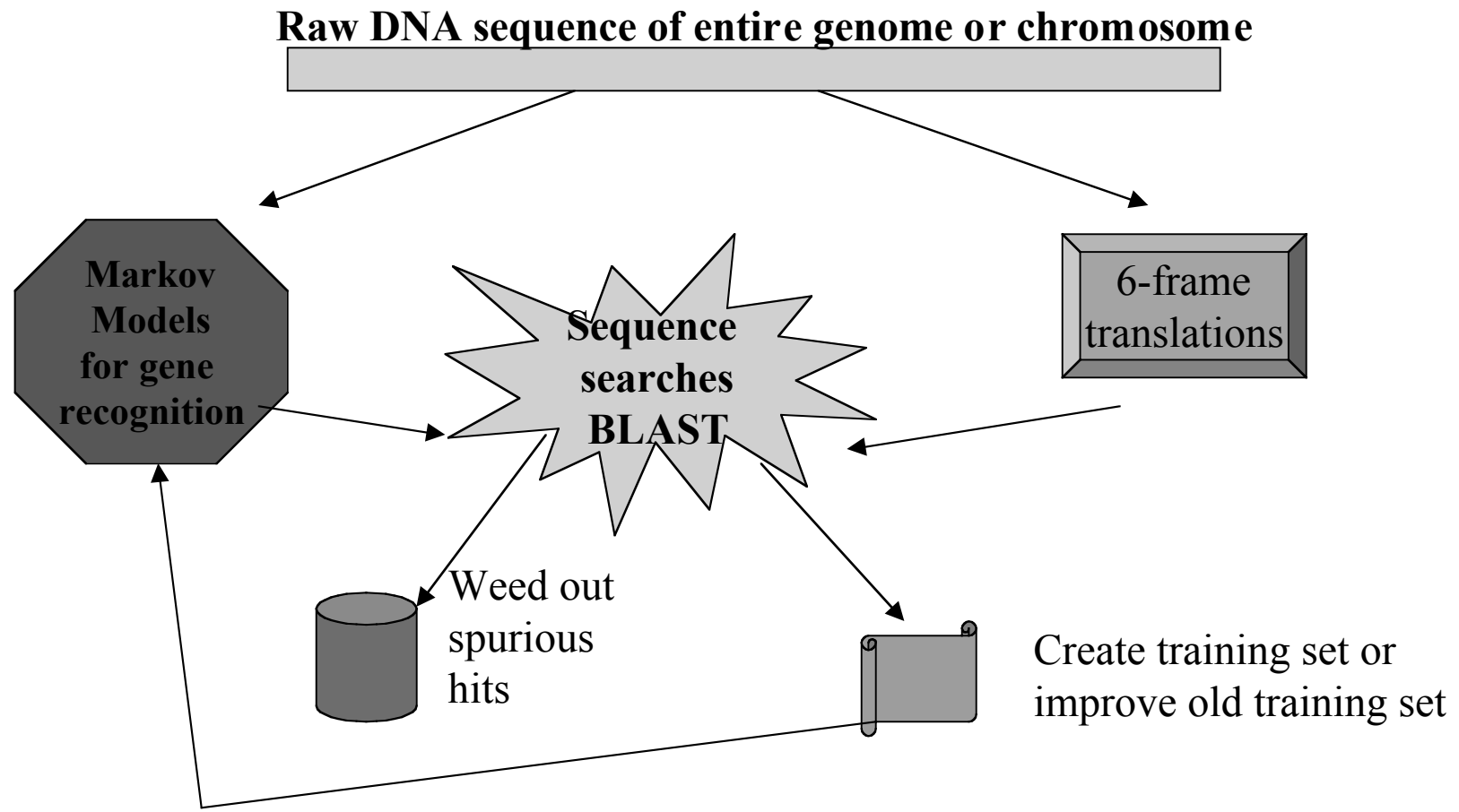
- Expression Arrays
- GFP fusion constructs
- The SAGE method
- Large scale transposon linked disruption

2) Computational approaches-

- Sequence similarity analysis of predicted proteins
- Nucleotide sequence analysis for regulatory elements and RNA genes
- Cross- genome evolutionary analysis to predict function

Sequence comparisons on the whole-genome scale

Stage 1: Gene recognition

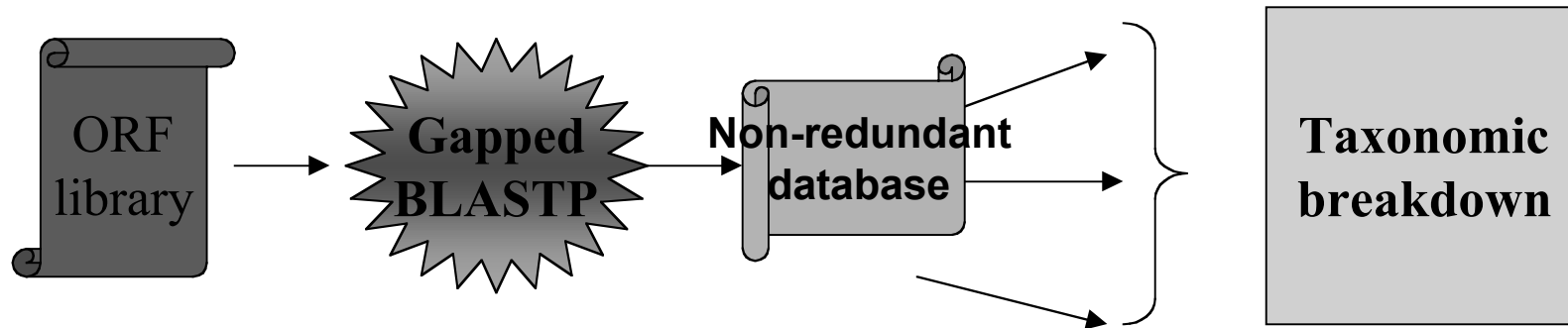


Sequence comparisons on a whole-genome scale

Stage 2: basic similarity searches and taxonomic breakdown

- Entirely automatic genome annotation is a dangerous myth
- ...but entirely manual analysis on genome scale is not feasible
- Current solution: semi-automatic genome analysis with manual intervention at the crucial steps

Primarily this involves batch searching of all ORFs (identified by the above procedures) using gapped BLAST and collecting the taxonomic distribution of best hits.



Done using the package SEALS: SEALS: a system for easy analysis of lots of sequences. Ismb. 1997;5:333-9. D. R. Walker and E. V. Koonin

Sequence comparisons on a whole-genome scale

Some caveats:

- *Statistical significance (Karlin-Altschul theory)*

- 1) **Highly significant hits - immediately relevant**
- 2) **Border-line significance- evaluate with more sensitive technique, use multiple lines of evidence**
- 3) **Low significance - discard.**

- *Filtering: Protein sequences are compositionally biased*

- 1) **Low complexity**
- 2) **Coiled coils**
- 3) **Membrane-spanning regions**

Detecting compositionally biased regions has two purposes:

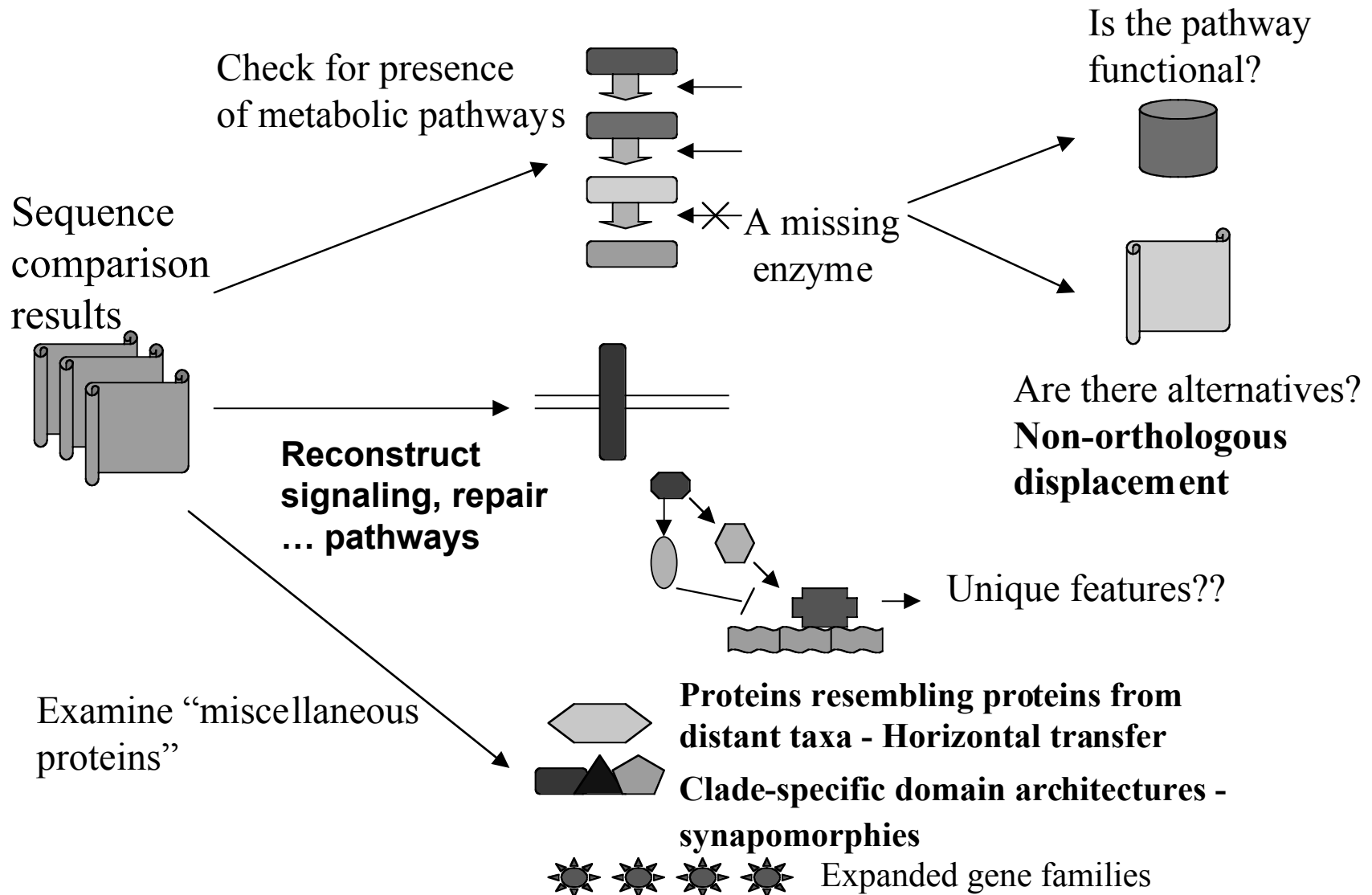
- 1) **mask these regions in sequence searches to prevent spurious hits**
- 2) **make functional predictions about these proteins**

Sequence comparisons on a whole-genome scale

Stage 3: “Deep” sequence analysis

- PSI-BLAST: Makes a profile of the alignments emerging in the first round of BLAST searches and proceeds to iterate with the profile till convergence**
- Profile analysis: Search the genome sequences with profiles for different domain families generated using PSI-BLAST**
- Structural analysis: Mapping sequence conservation to 3D structure in order to draw inferences on function.**
- Phylogenetic analysis: Carefully construct multiple alignments and make trees to explore evolutionary relationships.**

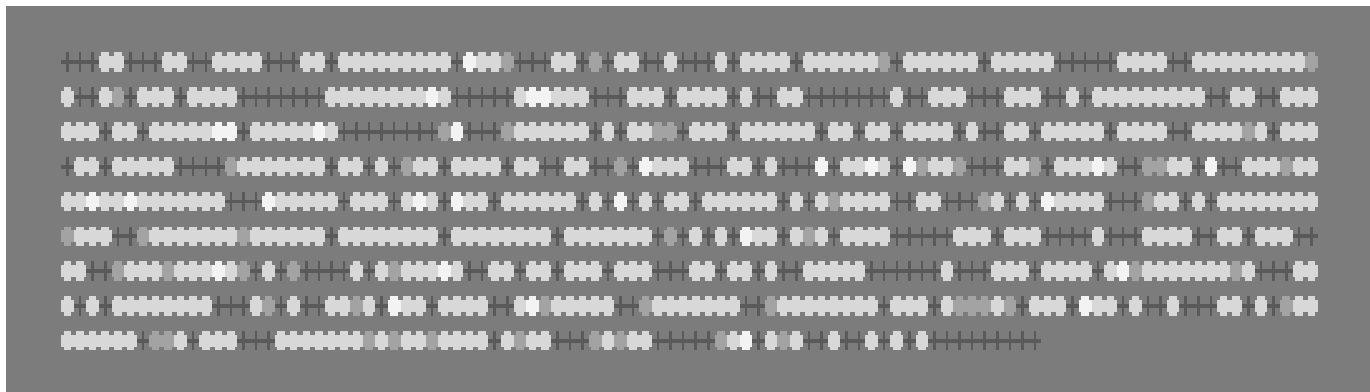
Finally: Reconstructing the organism



Chlamydia trachomatis: a case study

A leading causative agent of trachoma and genital tract infections.

An obligate intracellular form with 2 distinct “developmental” phases.



The genes in the Chlamydia genome are shown as circles whenever there is a significant hit to a protein in the database.

Turquoise - reliable best hits to other bacteria, magenta - eukaryotic hits, yellow - archaeal hits.

“Eukaryotic” genes in *Chlamydia trachomatis* - the result of

Horizontal gene transfer from the host

- **Eukaryotic chromatin-associated proteins:**

- **SET domain.**

- **The SWIB domain- 2 copies, one fused to topol**

Proposed functions: condensation of Chlamydial chromatin (unusual chromatin structure among bacteria)? Interaction with host chromatin?

- **Methionyl, Isoleucyl tRNA synthetases - translation.**

- **Glycogen metabolism enzymes, e.g. glycogen phosphorylase.**

- **ATP/ADP translocase- found only in plants, Chlamydia and Rickettsia.**

ATP scavenging from the host cell

- **At least 3 fatty acid metabolism enzymes distinctly related to plant homologs**

- **Thioredoxin peroxidase**

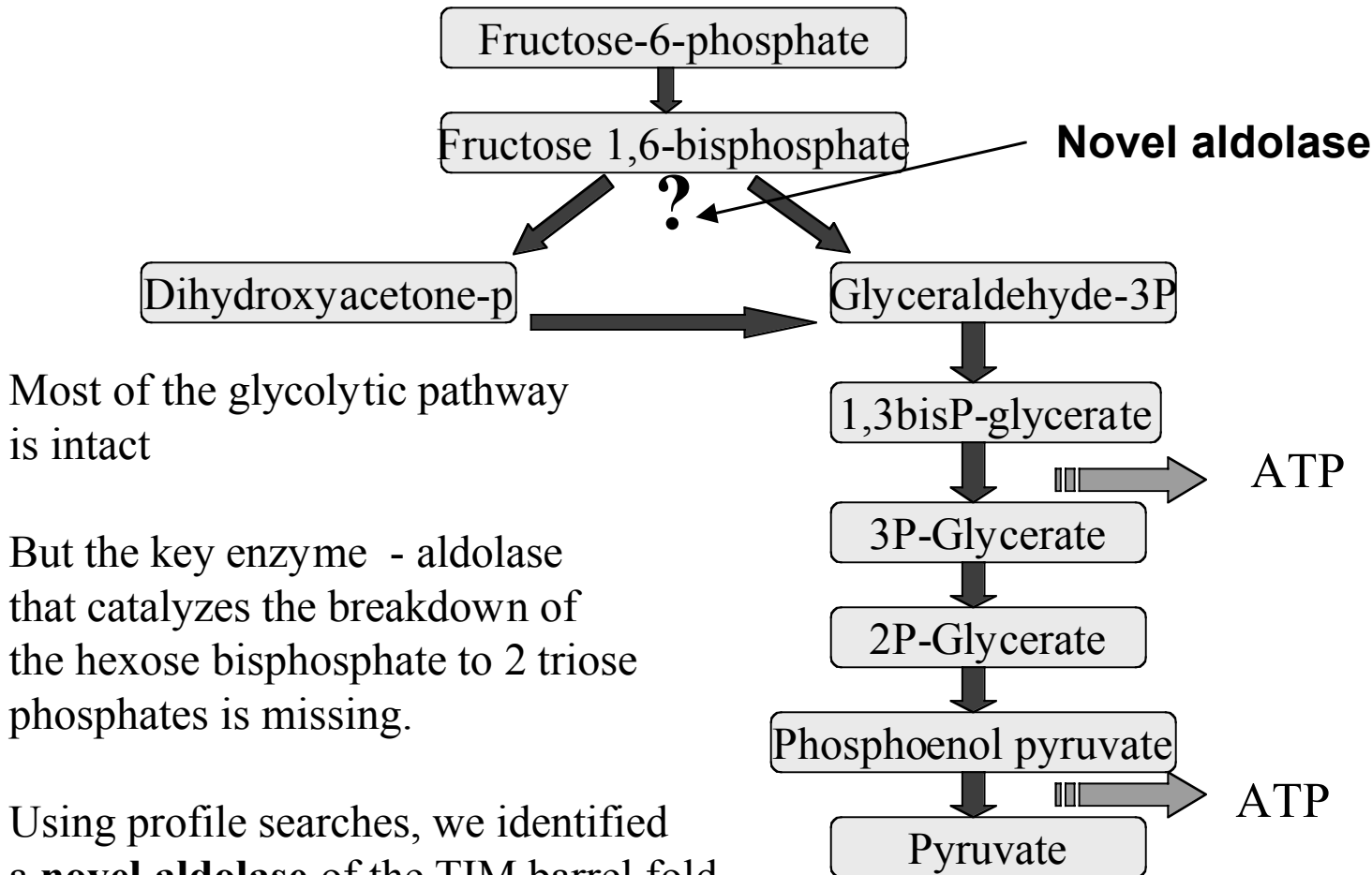
- **Superoxide dismutase**

Oxidative stress response in host cell

- **Secreted adenovirus-type thiol proteases**

Chlamydia trachomatis

Glycolysis and the case of the missing aldolase



Most of the glycolytic pathway is intact

But the key enzyme - aldolase that catalyzes the breakdown of the hexose bisphosphate to 2 triose phosphates is missing.

Using profile searches, we identified a **novel aldolase** of the TIM barrel fold- previously called dehydrin! **Subsequently, experimental evidence from *E coli* has confirmed this prediction.**

Unusual “bacterial” genes in Chlamydia

Na translocating NADH-ubiquinone
Oxidoreductase- Energy production
by way of sodium transport?

An expansion of phospholipases of the HKD superfamily- 6 members:

Type III secretion system to secrete virulence factors into host?

/ V-type ATPase
 operon-
 an adaptation to
 vacuolar
 environment?

Interferon gamma

Trp operon acquired from
proteobacteria- to counter inhibition by
Interferon gamma which limit cellular tryptohan

Two obligate intracellular pathogens- How do they compare?

We compared the 2 intracellular pathogens- *Chlamydia trachomatis* and *Rickettsia prowazekii*

As a result of intracellular life the pathogens have undergone **extensive gene loss**.

The pathogens have acquired **largely independent gene sets** from their hosts.

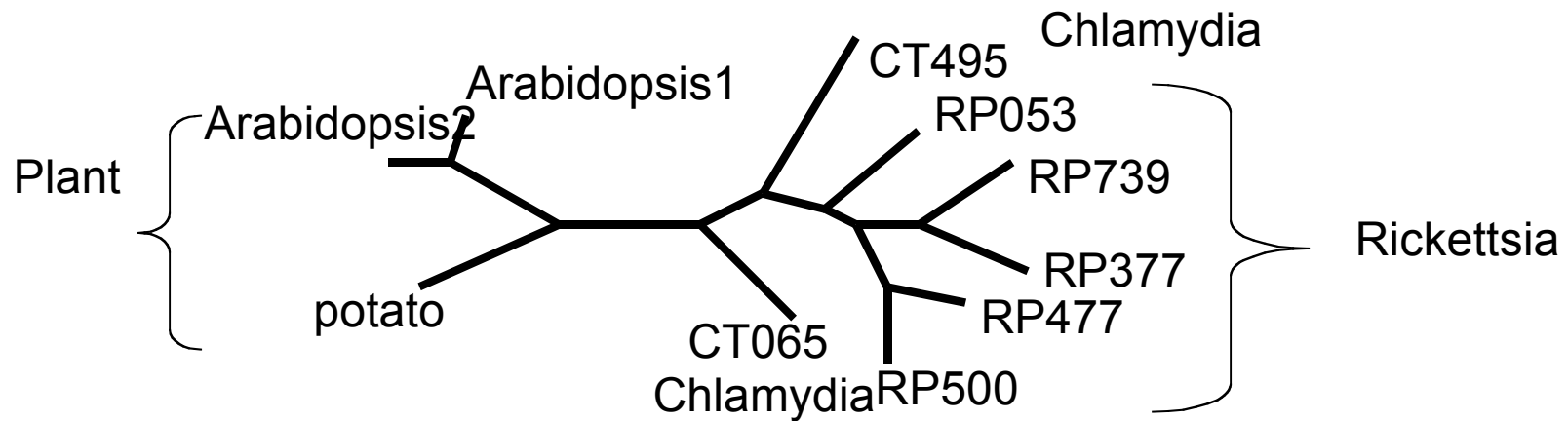
Reliable best hits to the 3 eukaryotic crown group taxa

Bacterium	Plants	Animals	Fungi	Total	
Chlamydia	16	8	2	26	
Rickettsia	0	11	4	15	

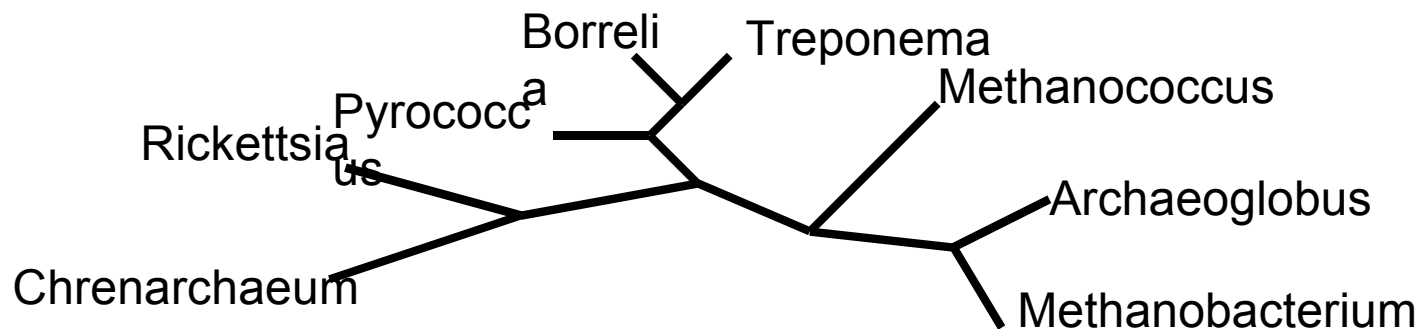
Chlamydia might have spent a substantial part of its evolutionary history in plant- related hosts (Acanthamoebae?) with subsequent transfer to an animal host. **Rickettsias** have been animal parasites for most of their history.

Gene Exchanges between intracellular resident prokaryotes

The **Adenine nucleotide translocators** responsible for “energy parasitism” of the Rickettsiae might have been acquired from the Chlamydiae as result of their shared environs.



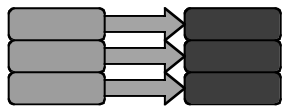
The unusual **lysyl tRNA synthetase** of Rickettsia has possibly been acquired from an intracellular symbiotic archaeon.



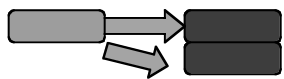
Comparing the two Spirochaetes

Treponema pallidum and *Borrelia burgdorferi*

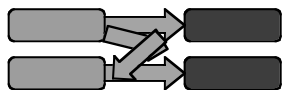
- The causative agents of syphilis and an ancient scourge of humans and Lyme's disease.
- Their genomes encode about 1000 genes each and offered the first case where one can estimate the evolutionary forces acting on 2 related average sized microbial genomes.
- To measure their relationship we used the **orthology coefficient**:
 - 1) Obtain counts of the orthologous genes in each of the genome.
 - 2) Then for each gene class which may be a super-family of proteins or a functional class such as transcription we compute the orthology coefficient (O) as below:



Complete 1:1 correspondence $O = 2 \cdot N_o / (N_b + N_t)$

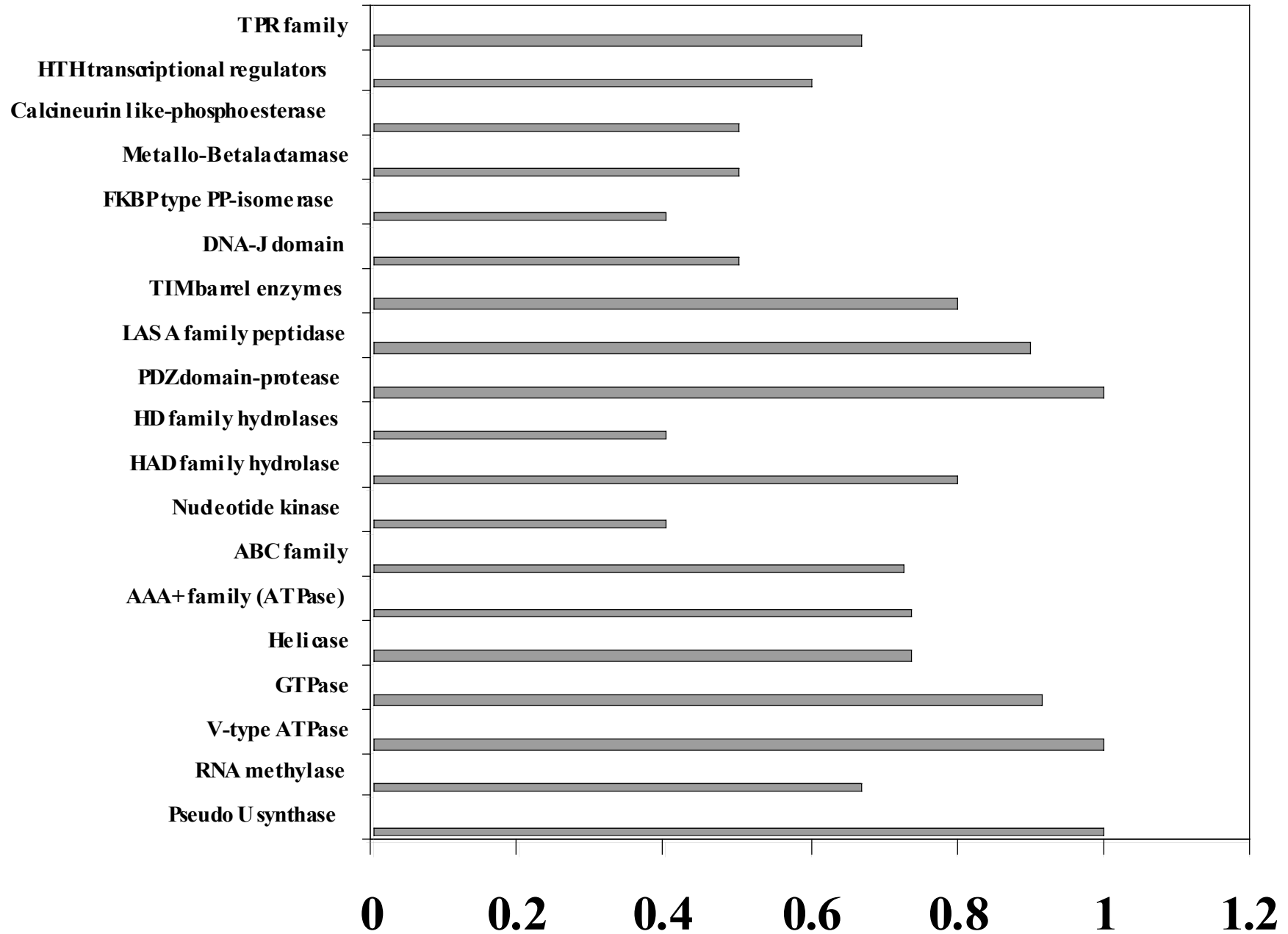


One-to-many or many-to-many correspondence

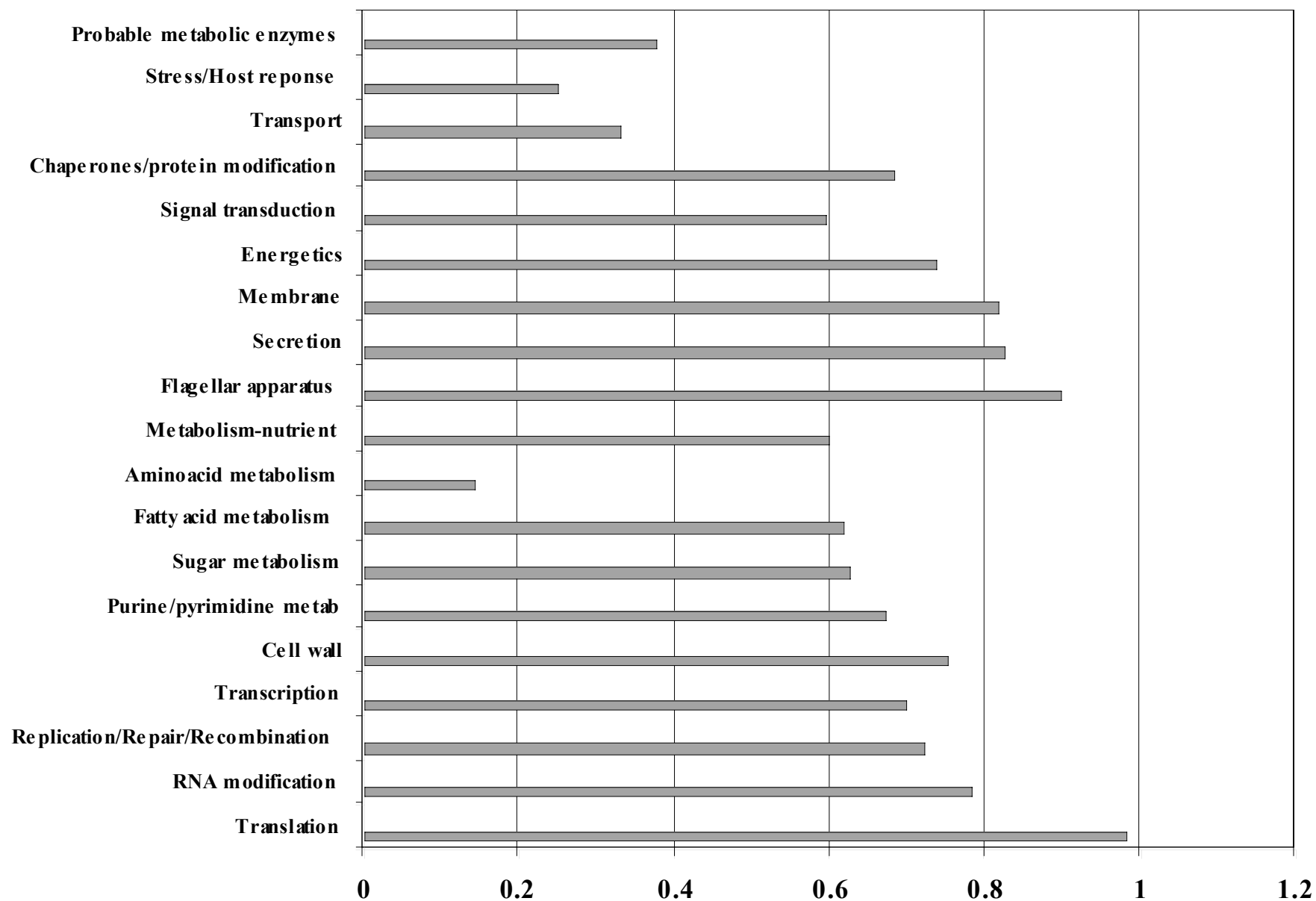


$$O = (N_{ob} + N_{ot}) / (N_b + N_t)$$

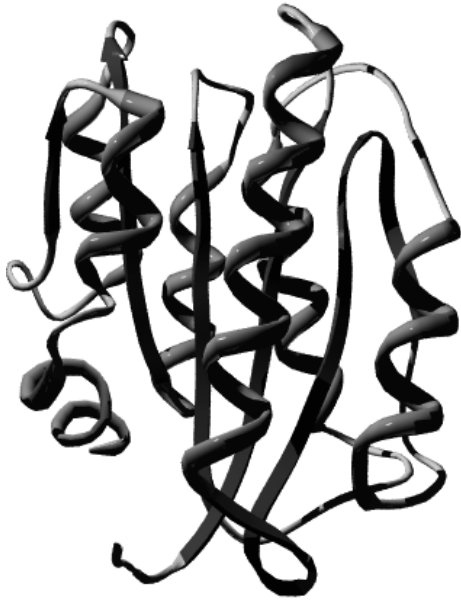
Orthology coefficient for gene families



Orthology Coefficient distribution in functional classes

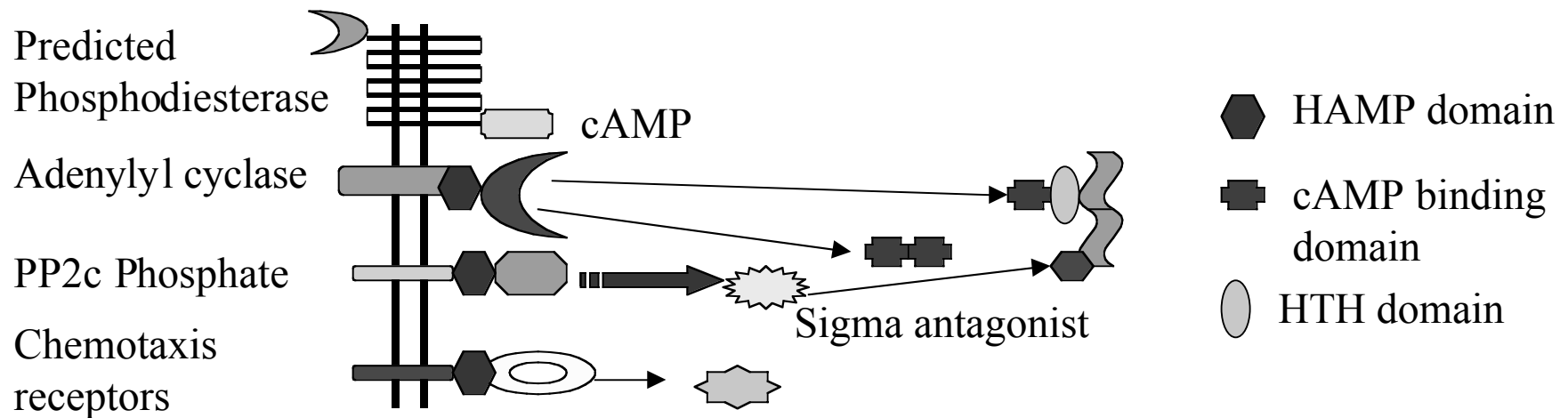


Novel findings in the Spirochaete genomes



- The presence of the Von Willebrand factor A domain-protein-protein interaction modules seen hitherto only in eukaryotes; in both *Borrelia* and *Treponema*.
- In *Borrelia* they are secreted and is likely to mediate interactions with host surface molecules similar to the Vwa Domains in the host integrin LFA-1 and Plasmodial TRAP protein.
- May have significance given the autoantibodies against LFA-1 in Lyme's disease.

Novel Multiple signaling pathways in *Treponema*





COG - Netscape


File Edit View Go Communicator Help

Bookmarks Location: <http://www.ncbi.nlm.nih.gov/COG/>

Instant Message Internet Lookup New&Cool

A Natural System of Gene Families from Complete Genomes



Clusters of Orthologous Groups (COGs) were delineated by comparing protein sequences encoded in 8 complete genomes, representing 6 major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.

[Science 1997 Oct 24;278\(5338\):631-637.](#)
[Curr Opin Struct Biol 1998 Jun;8\(3\):355-63](#)

Color	Code	Name	Genome size	Proteins
◆	E	<u><i>Escherichia coli</i></u>	4,653,831 bp	4283
◆	H	<u><i>Haemophilus influenzae</i></u>	1,830,240 bp	1703
◆	U	<u><i>Helicobacter pylori</i></u>	1,667,867 bp	1566 new
◆	G	<u><i>Mycoplasma genitalium</i></u>	580,073 bp	468
◆	P	<u><i>Mycoplasma pneumoniae</i></u>	816,394 bp	677
◆	C	Cyanobacteria - <u><i>Synechocystis</i></u>	3,573,470 bp	3168
◆	M	<u><i>Methanococcus jannaschii</i></u>	1,739,934 bp	1736
◆	Y	Yeast - <u><i>Saccharomyces cerevisiae</i></u>	12,068 Kbp	5932
Total				<u>19,533</u>

List of 864 COGs



Document: Done

COG - Netscape


File Edit View Go Communicator Help

Bookmarks Location: <http://www.ncbi.nlm.nih.gov/COG/>

Instant Message Internet Lookup New&Cool

A Natural System of Gene Families from Complete Genomes



Clusters of Orthologous Groups (COGs) were delineated by comparing protein sequences encoded in 8 complete genomes, representing 6 major phylogenetic lineages. Each COG consists of individual proteins or groups of paralogs from at least 3 lineages and thus corresponds to an ancient conserved domain.

[Science 1997 Oct 24;278\(5338\):631-637.](#)
[Curr Opin Struct Biol 1998 Jun;8\(3\):355-63](#)

Color	Code	Name	Genome size	Proteins
◆	E	<u><i>Escherichia coli</i></u>	4,653,831 bp	4283
◆	H	<u><i>Haemophilus influenzae</i></u>	1,830,240 bp	1703
◆	U	<u><i>Helicobacter pylori</i></u>	1,667,867 bp	1566 new
◆	G	<u><i>Mycoplasma genitalium</i></u>	580,073 bp	468
◆	P	<u><i>Mycoplasma pneumoniae</i></u>	816,394 bp	677
◆	C	Cyanobacteria - <u><i>Synechocystis</i></u>	3,573,470 bp	3168
◆	M	<u><i>Methanococcus jannaschii</i></u>	1,739,934 bp	1736
◆	Y	Yeast - <u><i>Saccharomyces cerevisiae</i></u>	12,068 Kbp	5932
Total				<u>19,533</u>

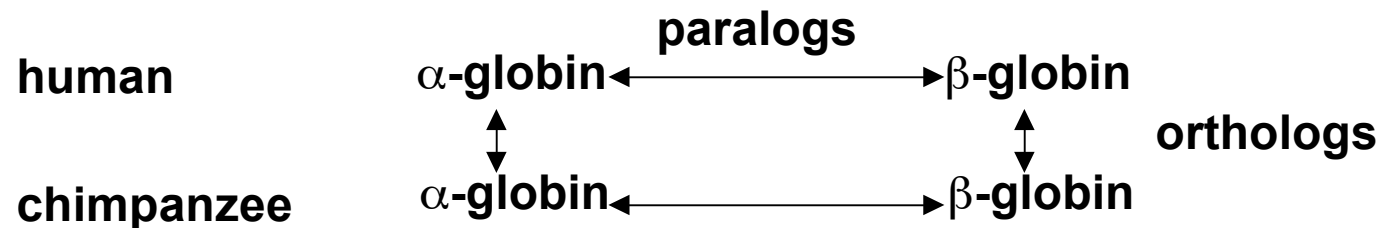
List of 864 COGs

Document: Done

Some definitions from evolutionary biology that are critical for genome comparison

Orthologs: gene in different species related by vertical descent

Paralogs: gene in the same species related by duplication



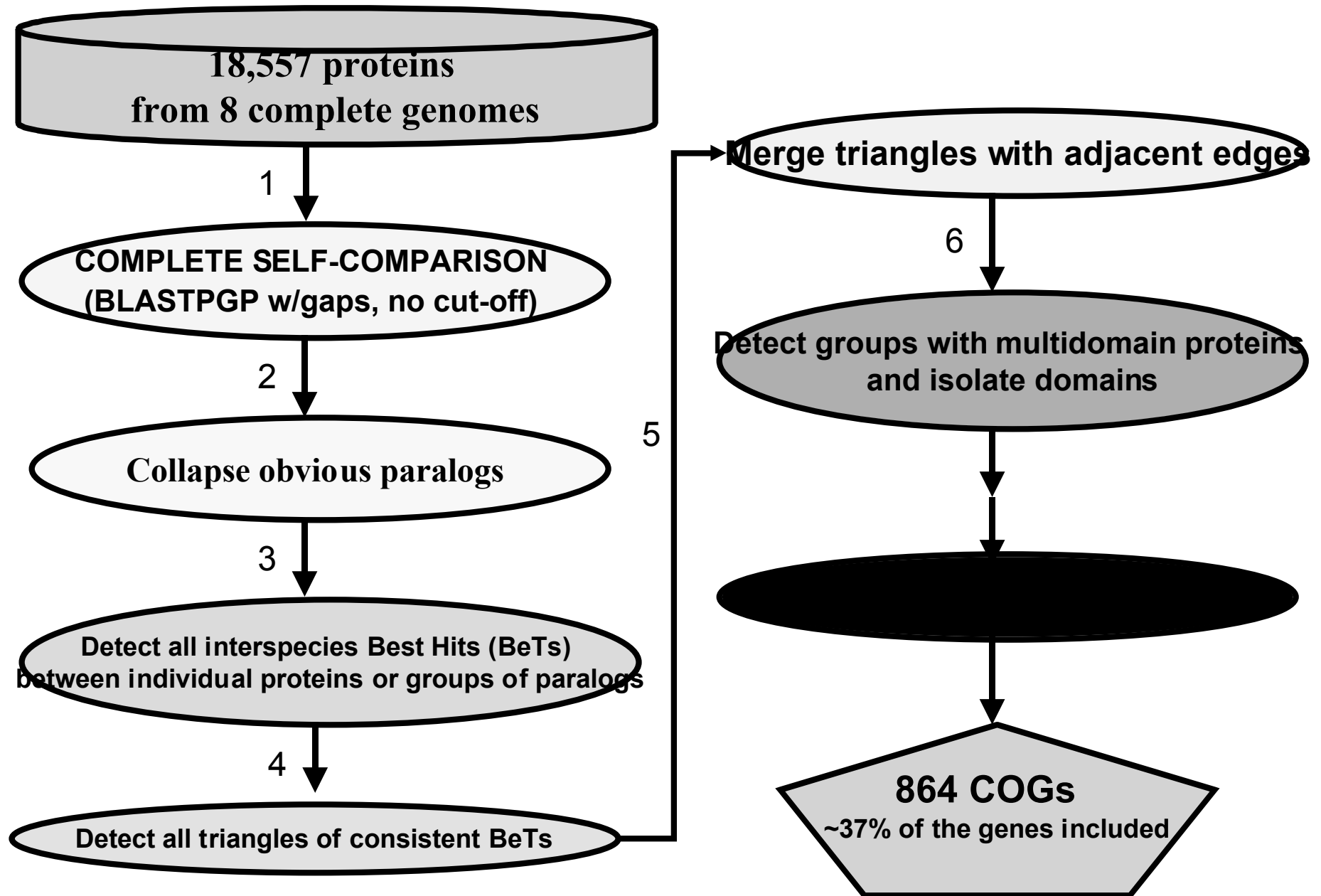
Goals:

- Using gene sets from complete genomes, delineate families of orthologs and paralogs - Clusters of Orthologous Groups (of genes) (COGs)
- Using COGs, develop an engine for functional annotation of new genomes
- Apply COGs for analysis of phylogenetic patterns

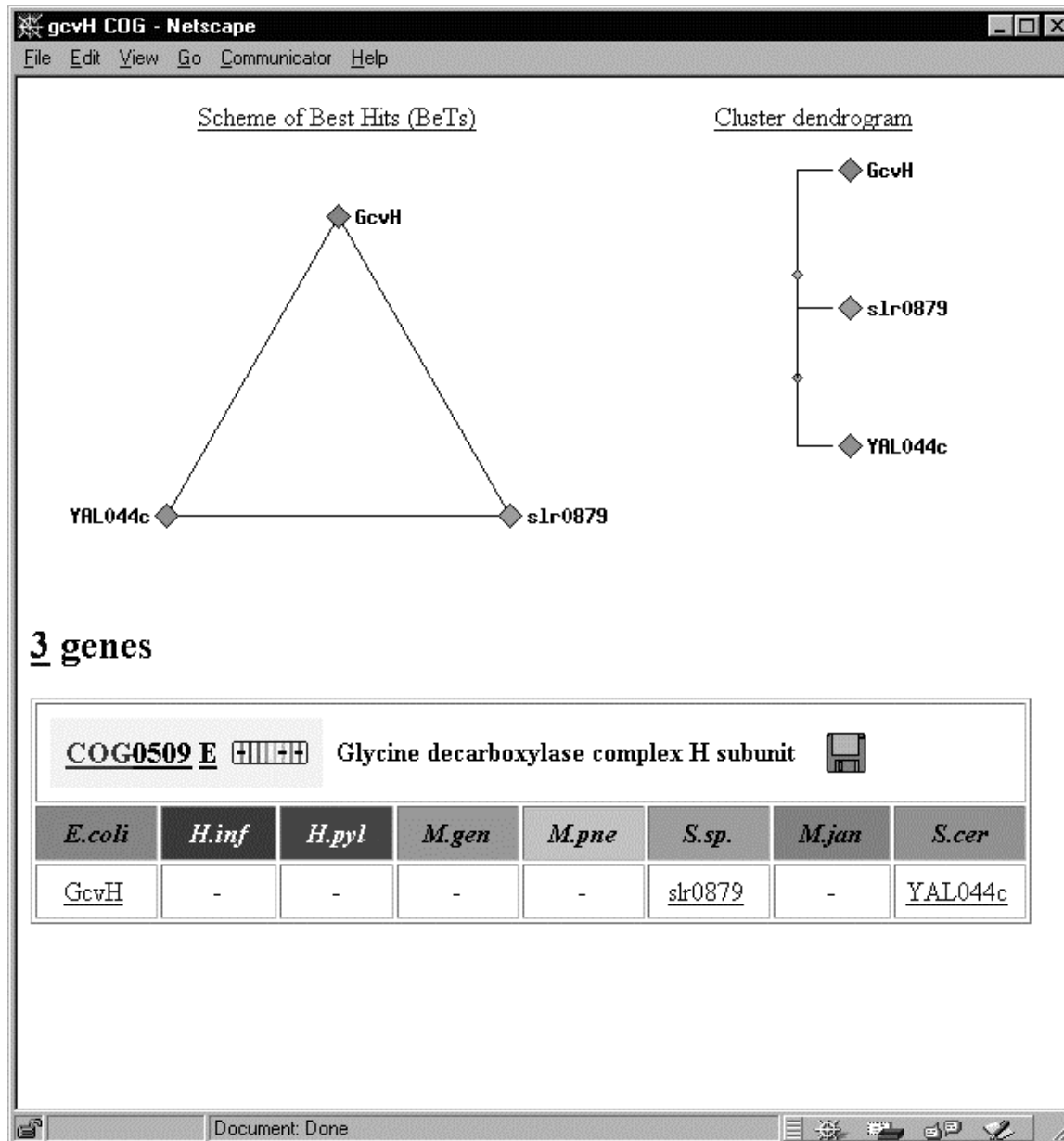
COG:

- group of homologous proteins such that all proteins from different species are orthologs (all proteins from the same species are paralogs)

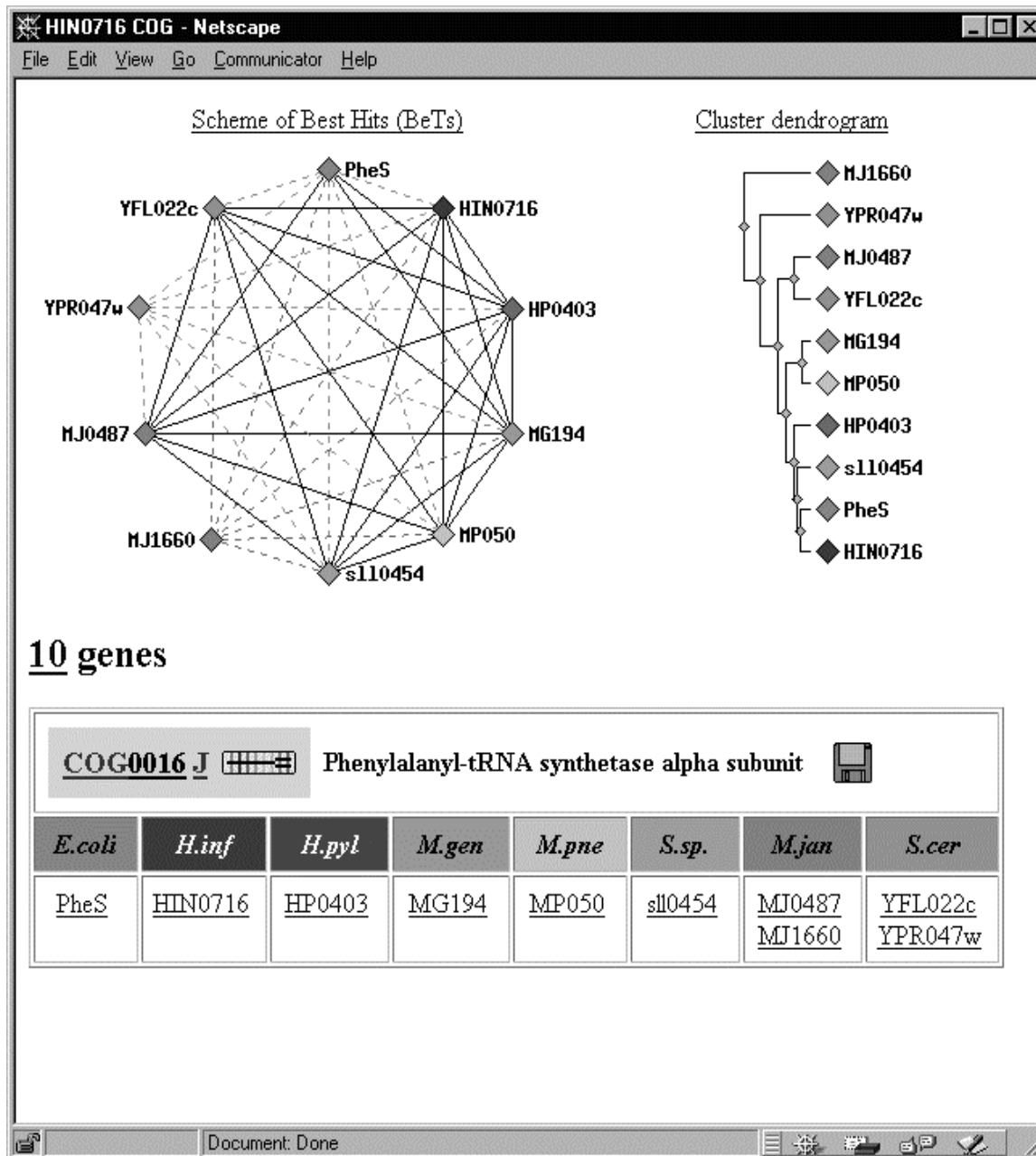
CONSTRUCTION OF COGs FOR 8 COMPLETE GENOMES

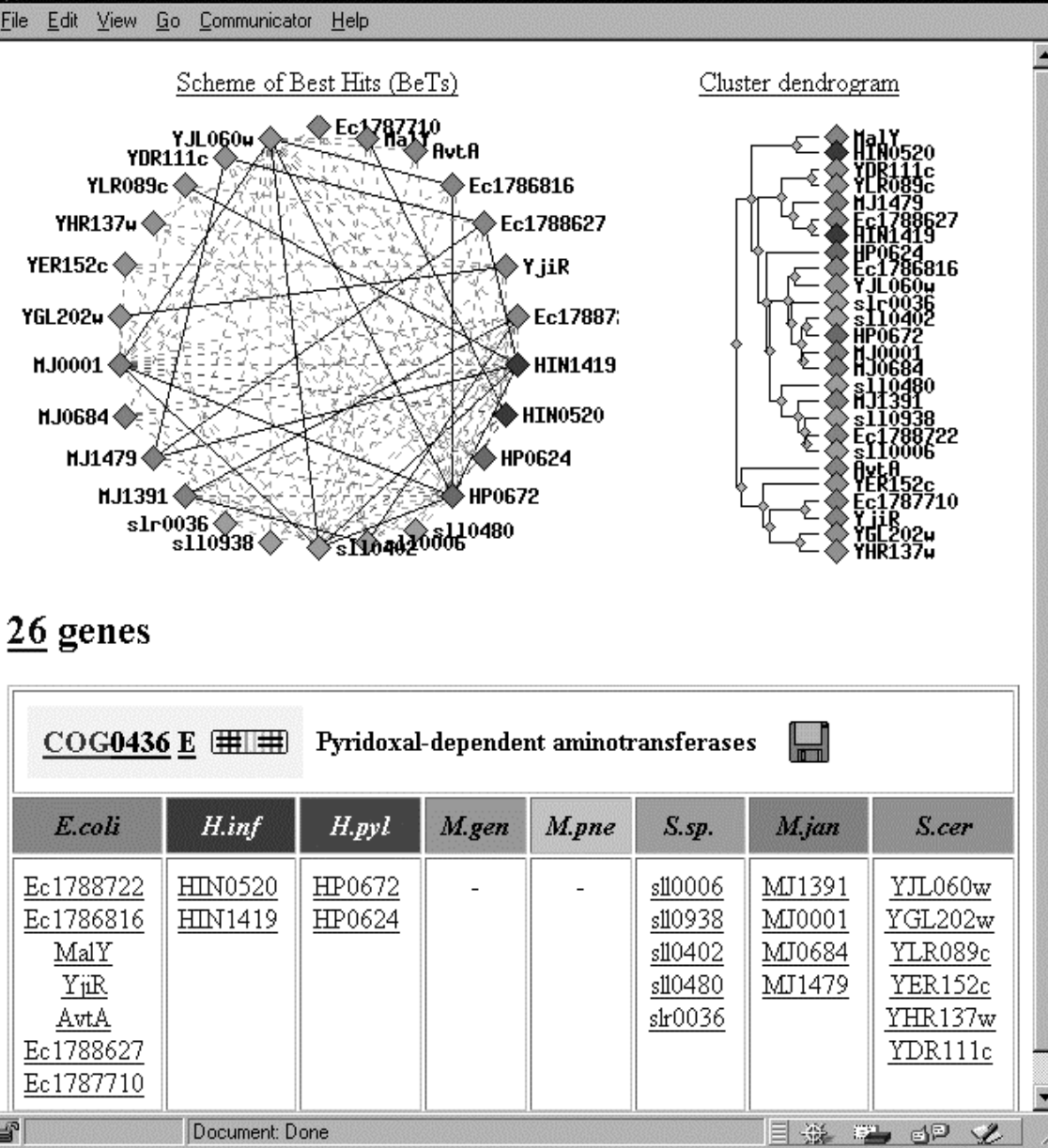
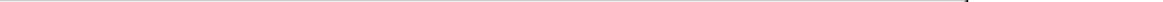


A TRIANGLE OF BeTs IS A MINIMAL, ELEMENTARY COG



A RELATIVELY SIMPLE COG PRODUCED BY MERGING ADJACENT TRIANGLES






CGO Functional Annotation Workbench

File Edit View Go Communicator Help

Back Forward Reload Home Search Guide Print Security Stop

Bookmarks Location: <http://www.ncbi.nlm.nih.gov/cgi-bin/COG/funu>

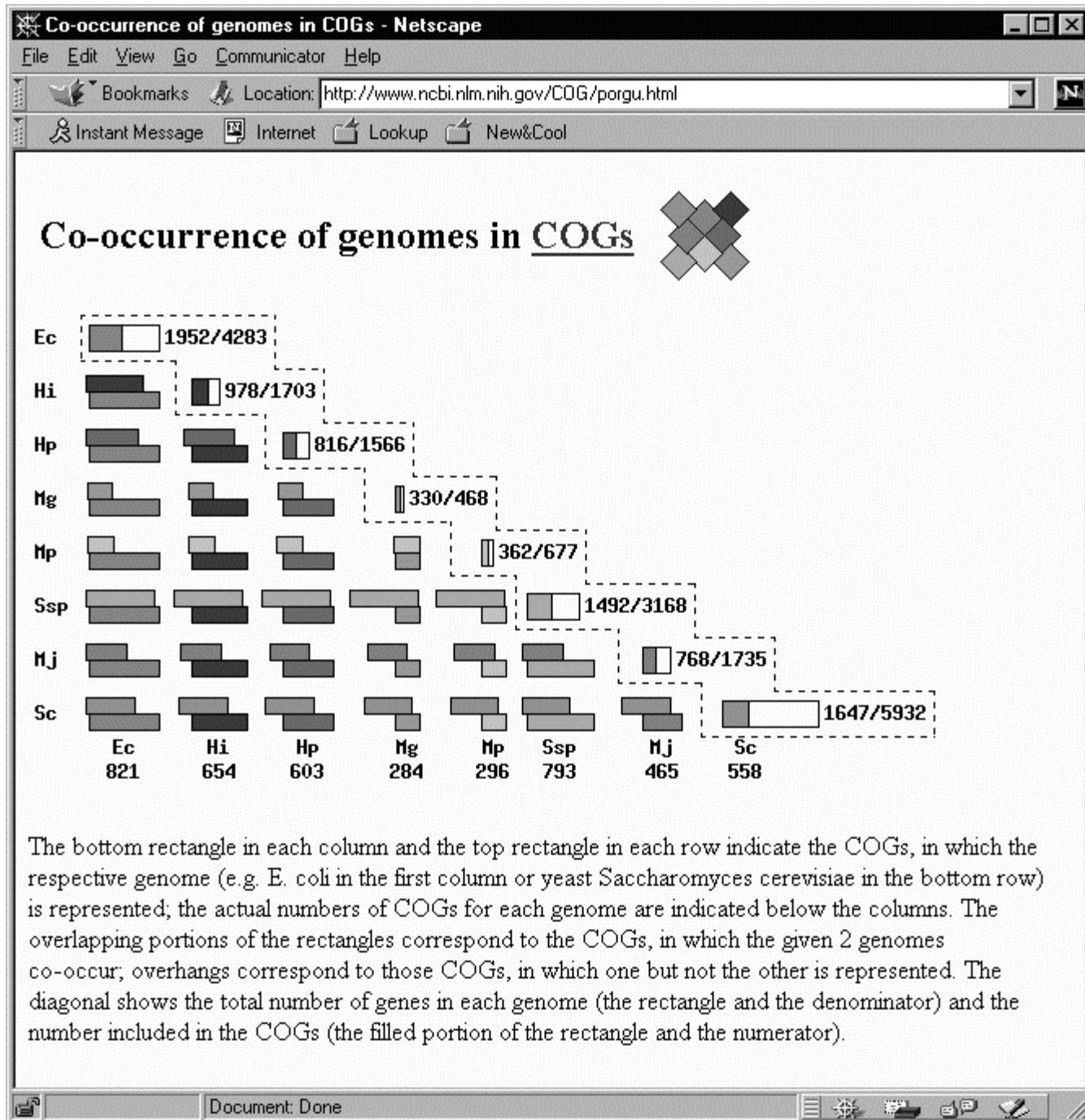
Instant Message Internet Lookup New&Cool



Functional annotation

Code	COGs	Domains	Description
Information storage and processing			
<u>J</u>	108	945	Translation, ribosomal structure and biogenesis
<u>K</u>	18	205	Transcription
<u>L</u>	64	654	Replication, repair, recombination
Cellular processes			
<u>O</u>	32	450	Molecular chaperones
<u>M</u>	47	409	Outer membrane, cell wall biogenesis
<u>N</u>	20	157	Secretion and motility
<u>P</u>	33	287	Inorganic ion transport and metabolism
Metabolism			
<u>C</u>	77	711	Energy production and conversion
<u>G</u>	33	301	Carbohydrate metabolism and transport
<u>-</u>			

Document: Done

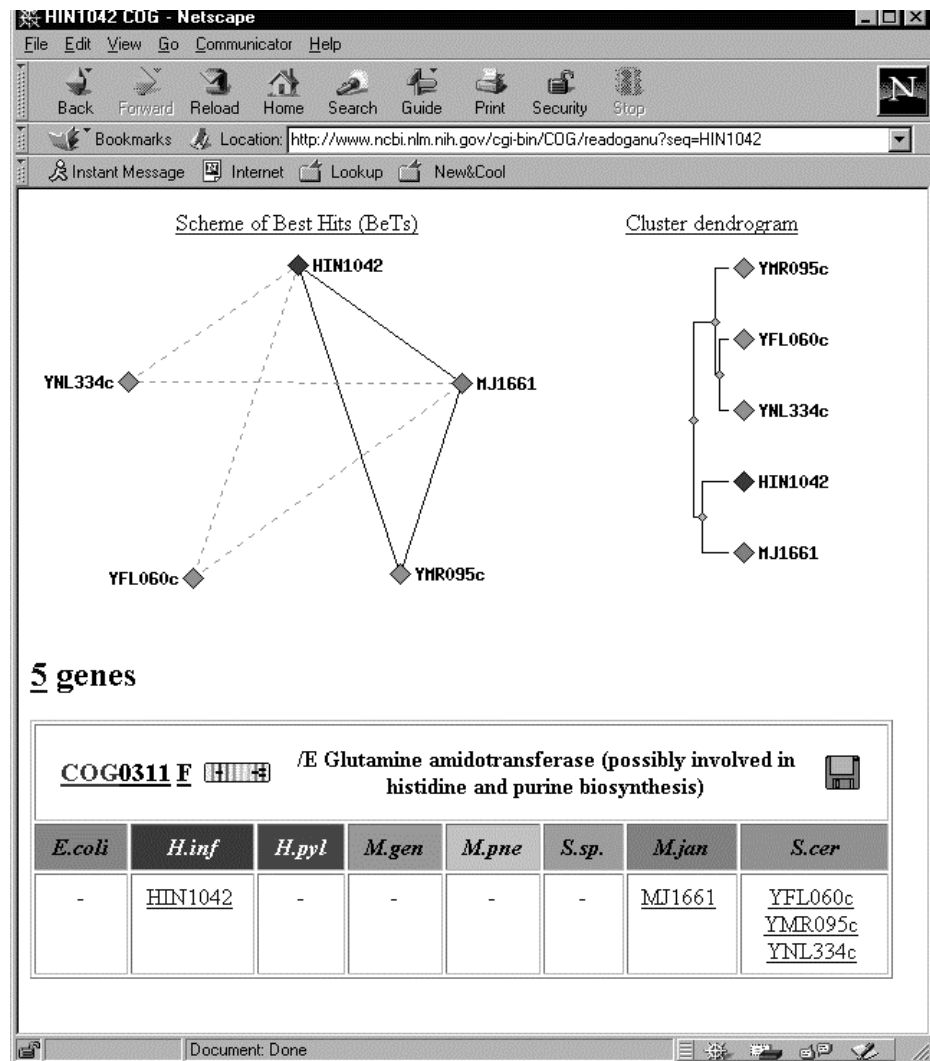
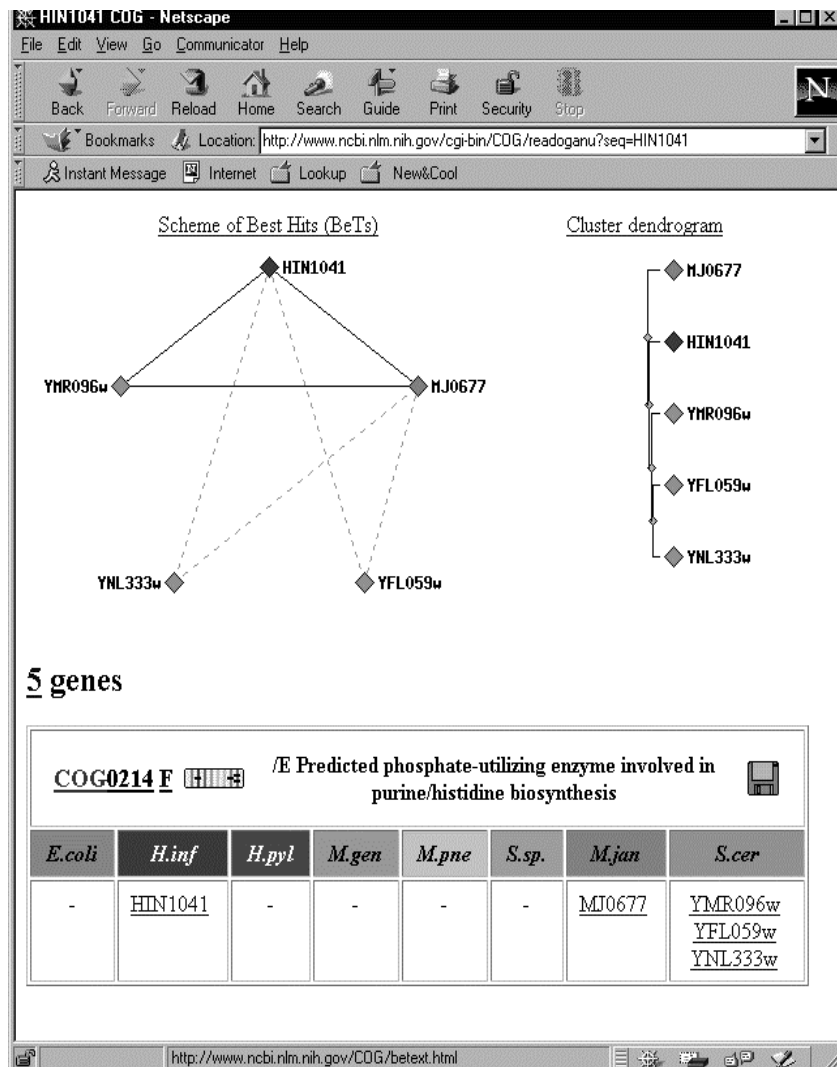


Bacteria + Eukarya + Archaea		Eukarya + Bacteria		Archaea + Bacteria		Bacteria only	
pattern	#	pattern	#	pattern	#	pattern	#
ehugpcmy	110	ehugpc-y	54	ehu--cm-	43	ehu--c--	77
ehu--cm-y	83	ehu--c-y	46	e----cm-	27	ehugpc--	41
eh---cm-y	39	e----c-y	41	e-u--cm-	16	e-u--c--	20
e----cm-y	21	eh---c-y	35	eh---cm-	13	eh-gpc--	13
e-u--cm-y	16	e-u--c-y	16	ehugpcm-	8	-hu--c--	8
ehu---my	11	eh-gpc-y	11	eh-gpcm-	7	e-ugpc--	3
-----cm-y	9	ehu---y	9	ehu---m-	4	ehugp---	2
eh----my	6	eh-gp--y	3	-h---cm-	3	eh--pc--	1
e-----my	6	e-u---y	2	e-u---m-	3	ehu-pc--	1
--u--cm-y	5	e--gpc-y	2	ehugp-m-	2	e--gpc--	1
eh-gpcmy	5	--u--c-y	2	e-ugpcm-	2	-hu-p---	1
-h----my	2	ehugp--y	2	e--gpcm-	2		
ehu-p-my	2	e-u-p--y	1	eh-gp-m-	1		
---gpcmy	2	-h---c-y	1	-hu---m-	1		
--ugpcmy	2	-hu---y	1	--u--cm-	1		
e-ugpcmy	2	e--gp--y	1	e--gp-m-	1		
e-u---my	1	e---p--y	1	ehu-p-m-	1		
ehugp-my	1	---gpc-y	1	--ugpcm-	1		
eh--pcmy	1	ehu-pc-y	1	eh--p-m-	1		
eh-gp-my	1	-h-gp--y	1	-hu-p-m-	1		
e--gp-my	1						
---gp-my	1						
22	327 (37%)	20	231 (26%)	20	138 (15%)	11	168 (19%)

Unique and rare phylogenetic patterns in COGs

-h---my

The only conserved metabolic pathway found in *H. influenzae* but NOT in *E. coli*



COGnitor...


Compare protein to COG database - Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Guide Print Security Stop

Bookmarks Location: http://www.ncbi.nlm.nih.gov/COG/cognitoru_MLH1.html

Instant Message Internet Lookup New&Cool

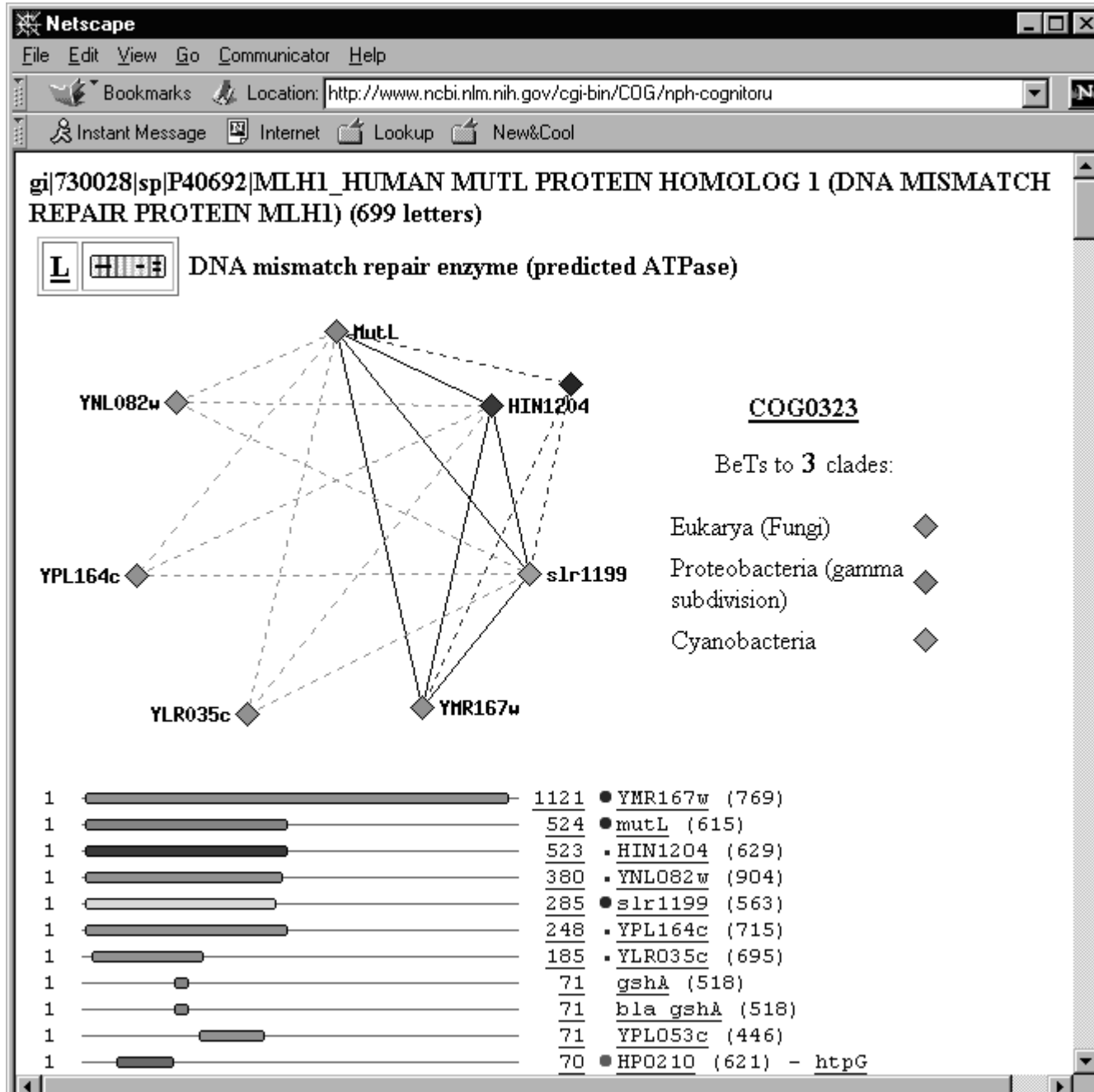
 **Compare MLH1 to COG database**

[Clear](#)

Please type your sequence and press a button.

```
>gi|730028|sp|P40692|MLH1_HUMAN MUTL PROTEIN HOMOLOG 1 (DNA MISMATCH RE
MSFVAGVIRRLDET VVNRIAAGEVIQRPANA IKEMIENCLDAKSTSIQVIVKEGGLKLIQIQDNGTGIRK
EDLDIVCERFTTSKLQSFEDLASISTYGFGEALASISHVAHVTTITTKTADGKCAYRASYS DGKLKAPPK
PCAGNQGTQITVEDLFYNIATRRKALKNPSE EYGKILEVVGRYSVHNAGISFSVKKQGETVADVRTL PNA
STVDNIRSIFGNAVSRELIEIGCEDKTLAFKMNGYISNANYSVKKCIFLLFINHRLVESTSLRKA IETVY
AAYLPKNTHPFLYLSLEISPQNV DVNVHPTKHEVHFLHEESILERVQQHIESKLLGSNSSRMYFTQTLLP
GLAGPSGEMVKSTTSLTSSSTSGSSDKVYAHQMVRTDSREQKLDAFLQPLSKPLSSQPQAIVTEDKTD IS
SGRARQQDEEMLELPAPAEVAAKNQSLE GDTTKGTSEMSEKRGPTSSNPRKRHRESDVEMVEDDSRKEM
TAACTPRRRIINLTSVLSLQEEINEQGHEVLREMLHNHSFVGC VNPQWALA QHQTKLYLLNTTKLSEELF
YQIL IYDFANFGVLRRLSEPA PLFDLAMLALDSPESGWTEEDGPKEGLAEYIVEFLKKKAEM LADYFSLEI
DEEGNLIGLPLLIDNYVPPLEGLPIFILRLATEVNWDEEKECFESLSKECAMFY SIRQYISEESTLSG
```

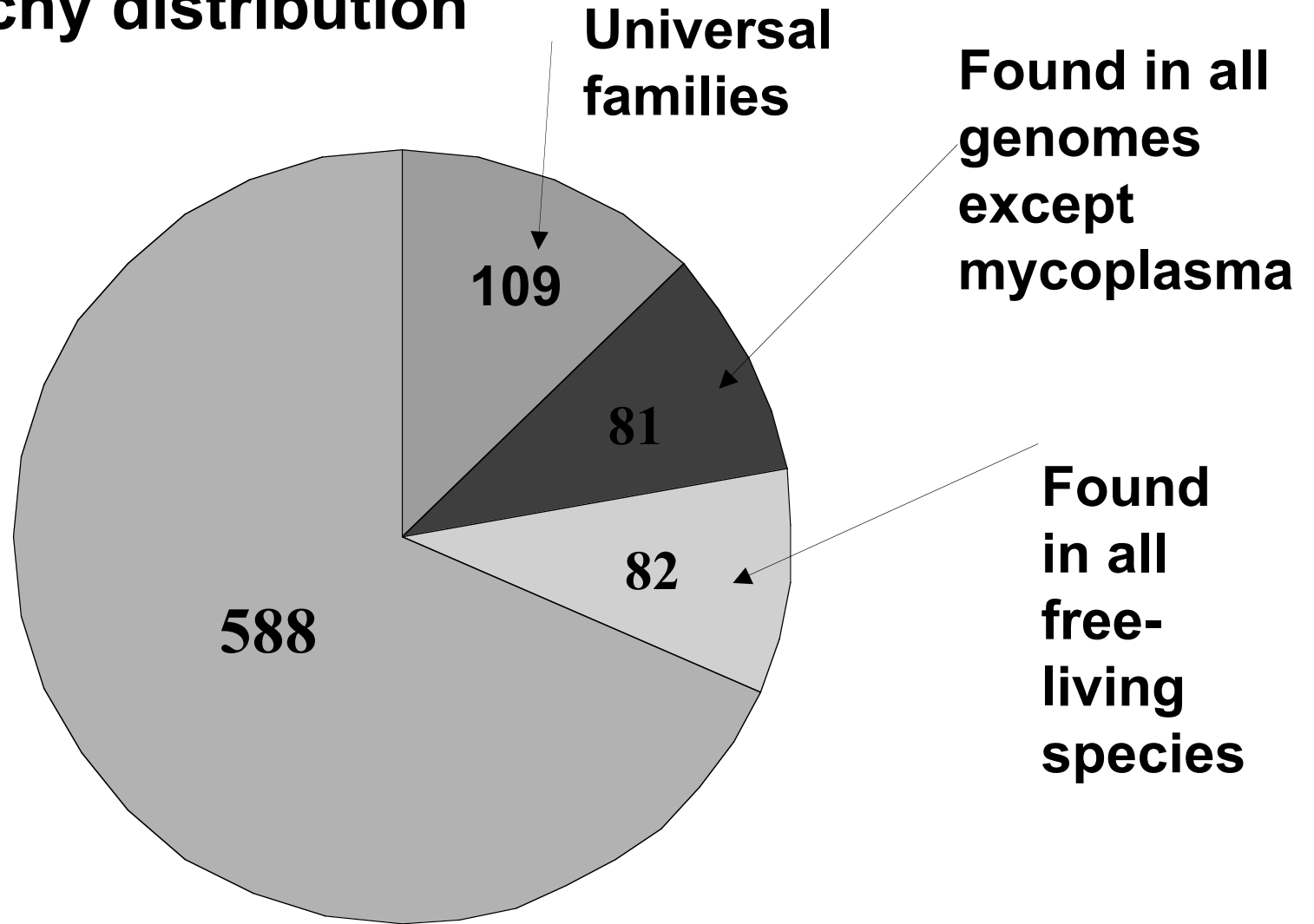
...IN ACTION



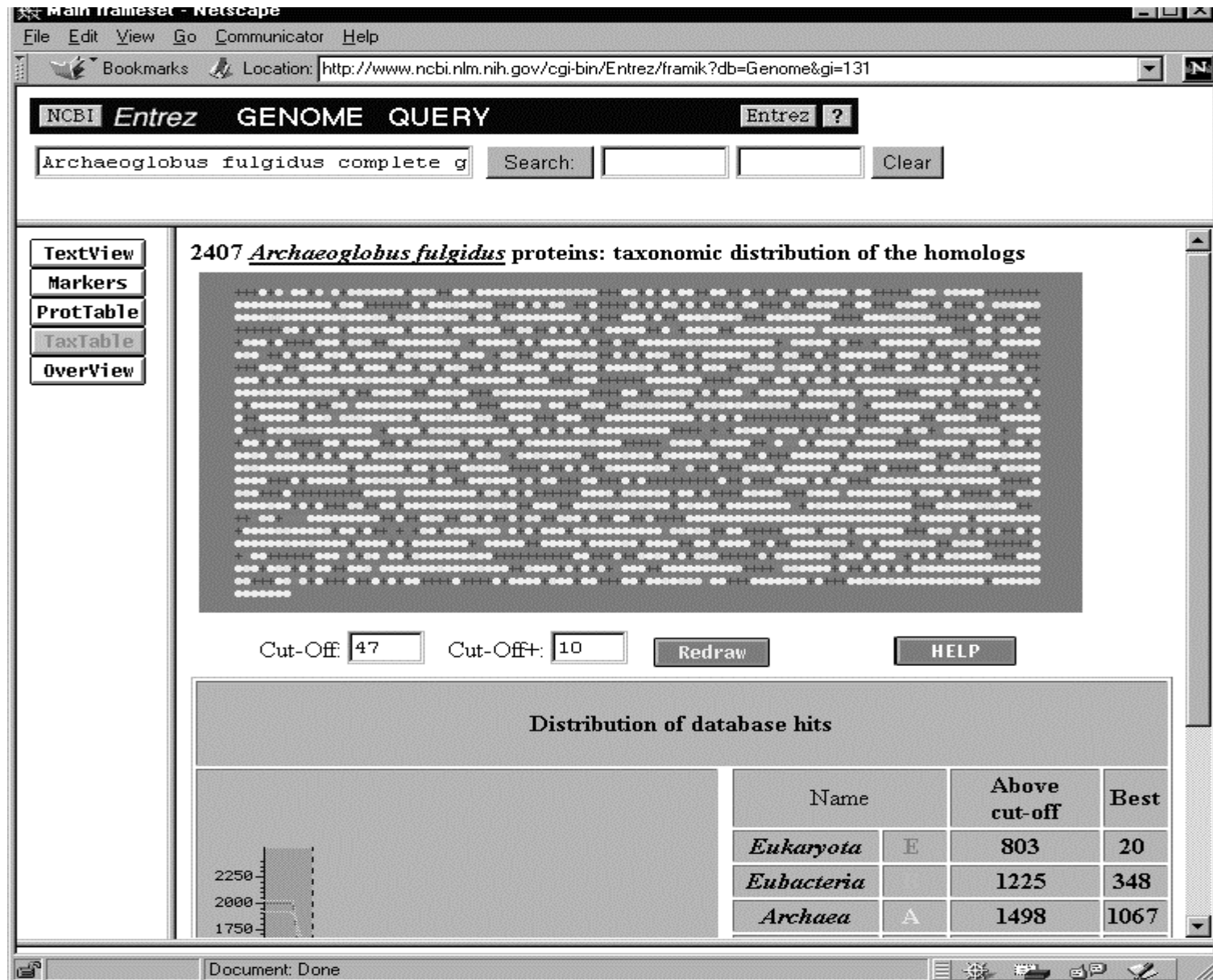
REPRESENTATION OF GENOMES IN COGs

<i>Escherichia coli</i>	4,289	2,003 (47%)
<i>Haemophilus influenzae</i>	1,717	979 (57%)
<i>Helicobacter pylori</i>	1,566	841 (54%)
<i>Synechocystis sp.</i>	3,169	1,551 (49%)
<i>Borellia burgdorferi</i>	850	483 (57%)
<i>Bacillus subtilis</i>	4,100	1,945 (47%)
<i>Mycoplasma genitalium</i>	467	341 (75%)
<i>Mycoplasma pneumoniae</i>	677	378 (56%)
<i>Methanococcus jannaschii</i>	1,715	830 (48%)
<i>Methanobacterium thermoautotrophicum</i>	1,869	897 (48%)
<i>Archaeoglobus fulgidus</i>	2,407	1,131 (47%)
<i>Saccharomyces cerevisiae</i>	5,932	1,551 (26%)
<i>Caenorhabditis elegans</i>	12,178	2,172 (18%)

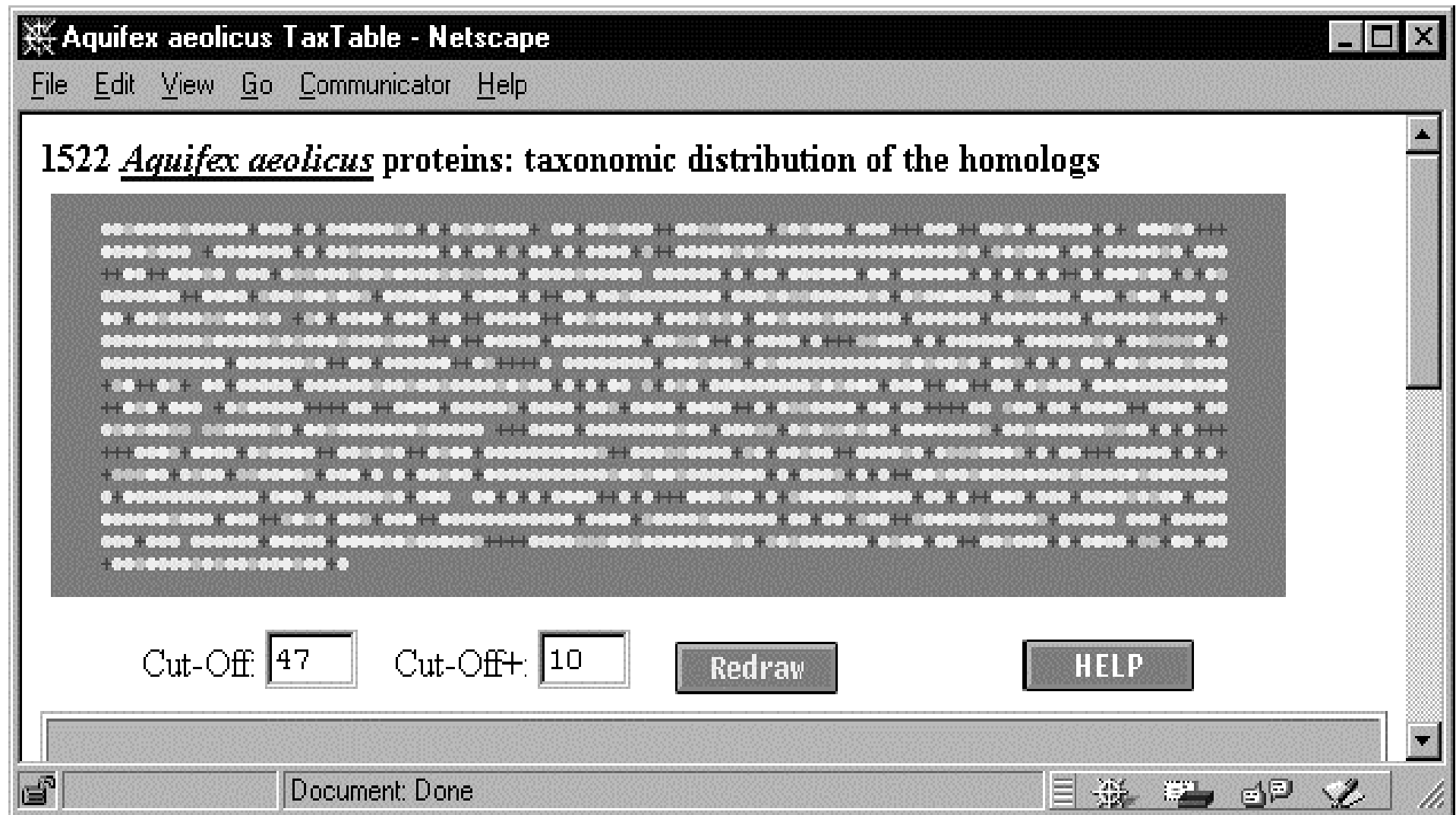
Universal gene families and families with patchy distribution



Significant horizontal gene transfer from bacteria to archaea



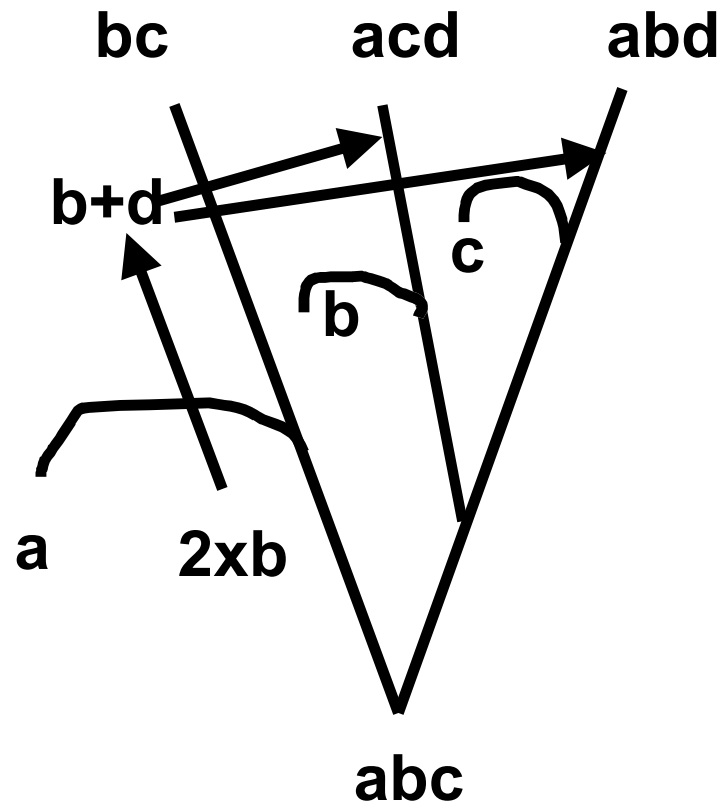
Massive horizontal gene transfer from archaea to a hyperthermophilic bacterium



Significant horizontal gene transfer from eukaryotes to a (facultative) symbiotic bacterium



Lineage-specific gene loss and horizontal gene exchange - formative forces of evolution



Lineage-specific gene loss, horizontal gene transfer

↓
Few universal families

↓
Broad distribution of
phylogenetic patterns

**Non-orthologous gene displacement:
same (essential) function - unrelated proteins**

**Not among the universal families -
and the strongest case of
non-orthologous gene displacement:
the basic DNA replication machinery**

The principal enzymes of DNA replication in bacteria and eukarya/archaea seem to have evolved independently

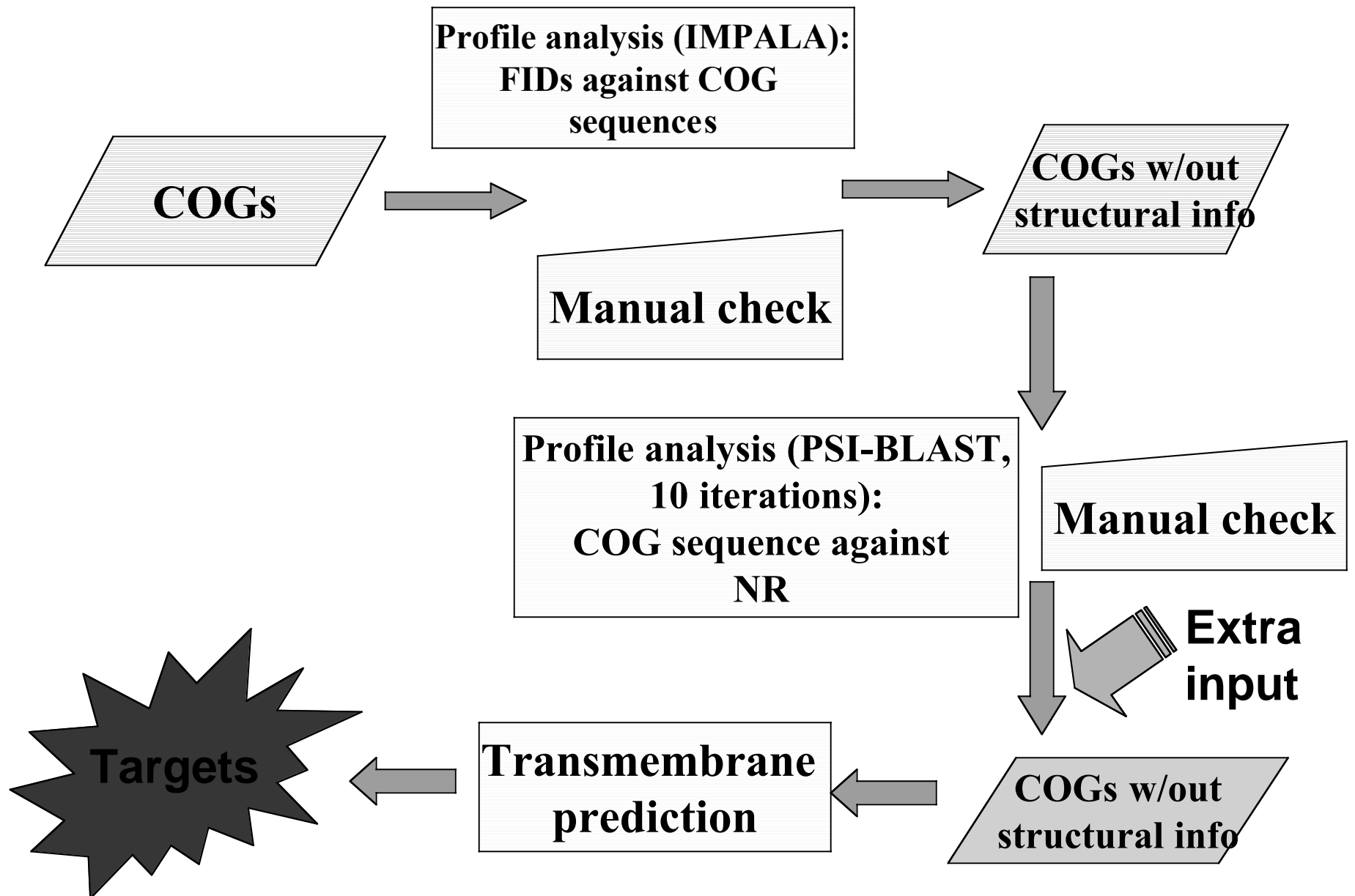
Replication enzyme	Bacteria	Archaea/ Eukarya	Relationship
DNA polymerase	PolC	PolB	Unrelated
Principal helicase	DnaB	SFI/II DNA helicases	Distantly related Independent origin
Primase	DnaG	PRIM1/2	Unrelated
Principal topo-isomerase	Topo IA	Topo IB	Unrelated
Initiator ATPase	DnaA	MCM family	Distantly related Independent origin

STRUCTURAL GENOMICS:

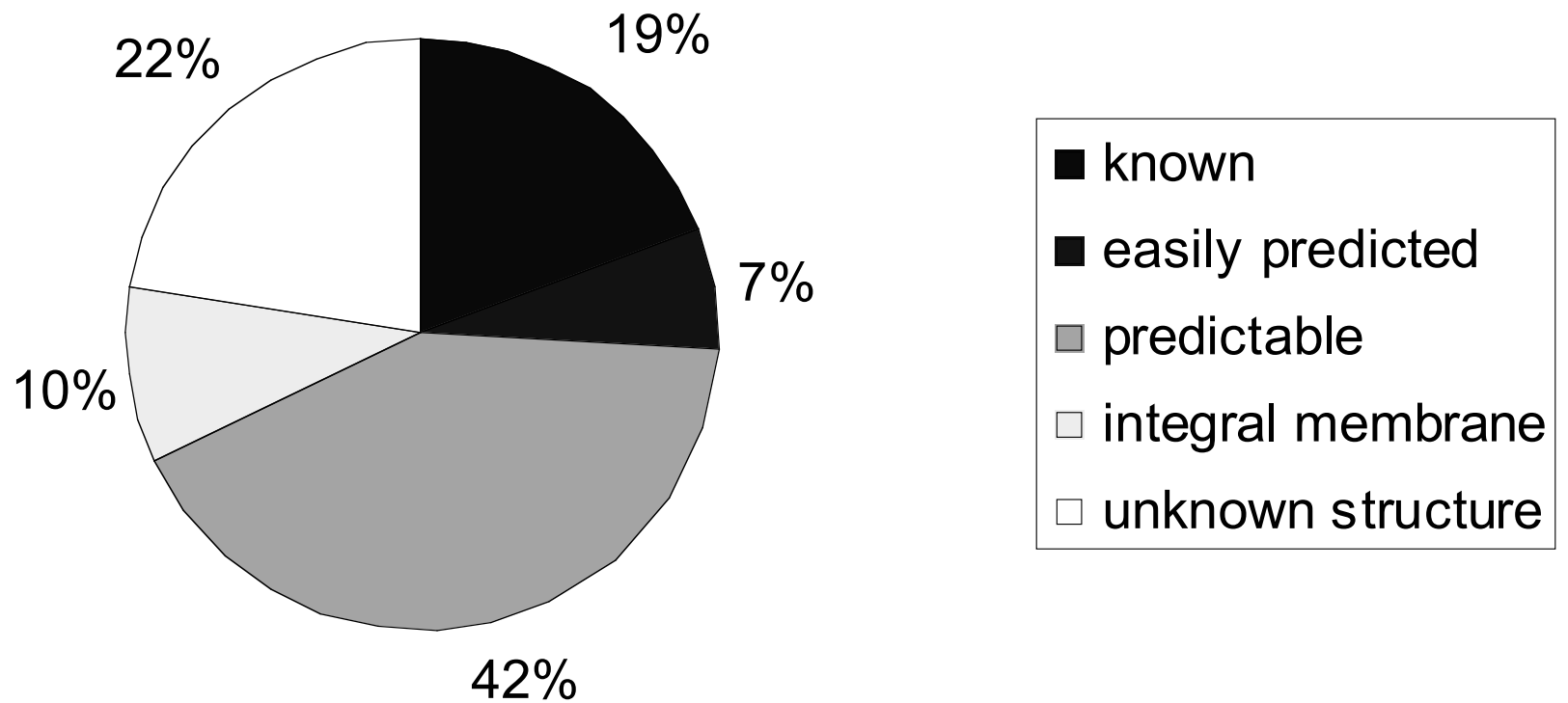
Systematic determination of protein 3D structures that

- are predicted to represent new folds or significant variants of known folds**
- correspond to highly conserved protein families or families with “interesting” phylogenetic patterns (e.g. specific for pathogens)**

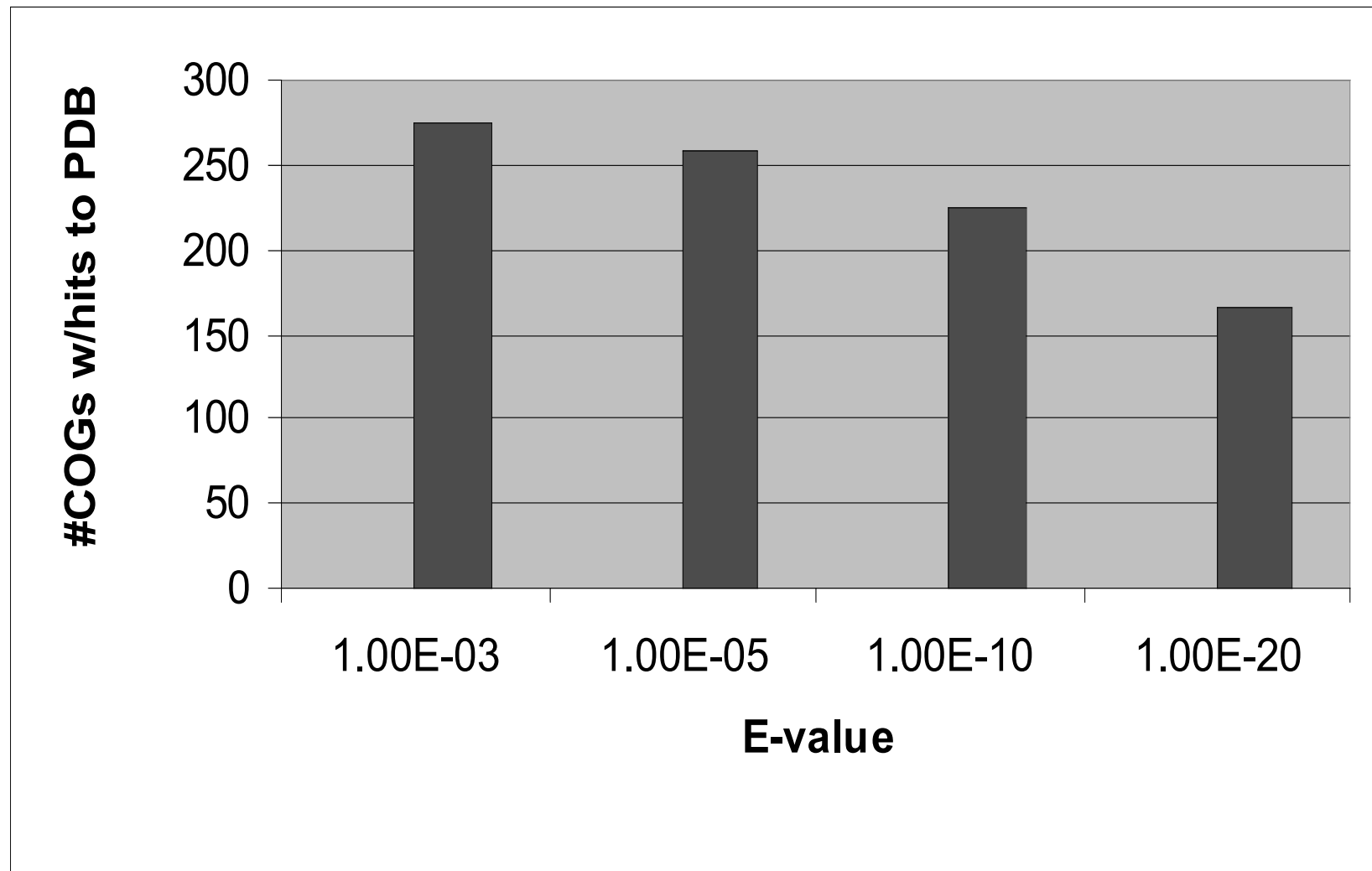
Identification/prediction of folds in COGs



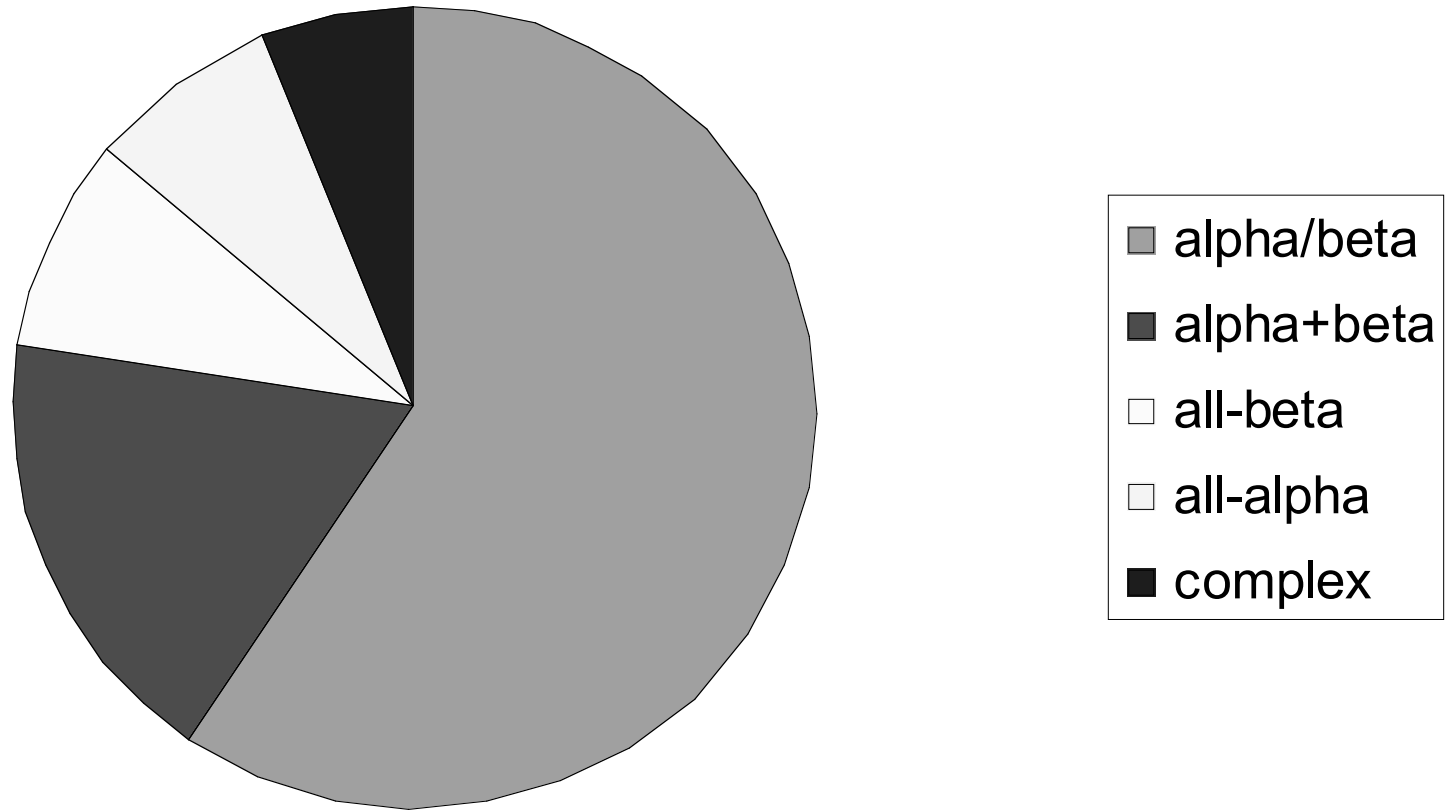
Structure and structure prediction in 864 COGs



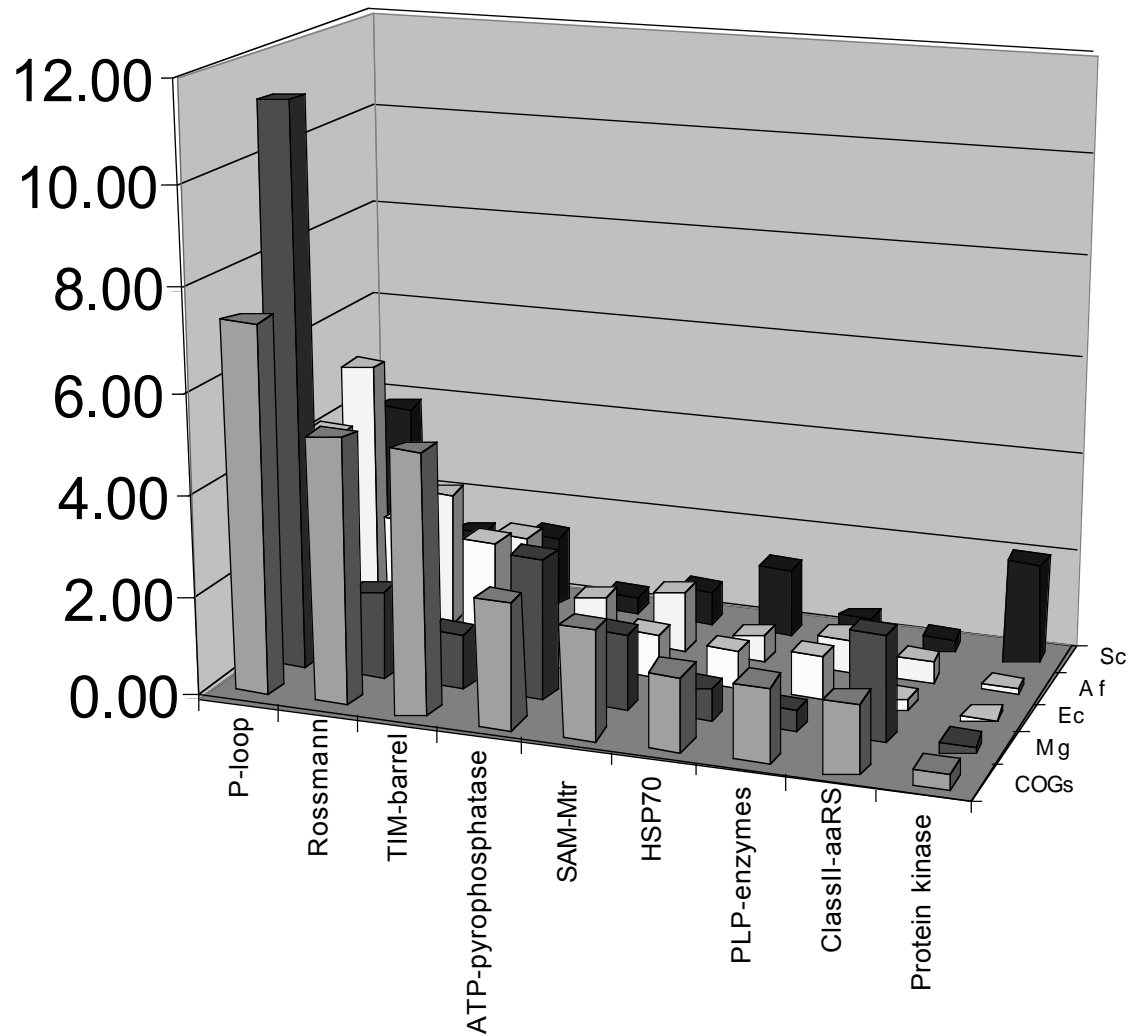
Fold prediction for COGs by BLAST alone



Protein structural classes in the COGs



Top folds in COGs and genomes



CONCLUSIONS

WHEN APPLIED WITHIN A COHERENT STRATEGY, CURRENT METHODS FOR PROTEIN SEQUENCE AND STRUCTURE ANALYSIS ALLOW US TO EXTRACT FROM GENOME SEQUENCES A WEALTH OF INFORMATION FOR FUNCTIONAL AND EVOLUTIONARY INFERENCES

THE NEW VIEW OF GENOME EVOLUTION:

- MAJOR ROLE OF HORIZONTAL GENE TRANSFER AND CLADE-SPECIFIC GENE LOSS IN EVOLUTION**
- ASYNCHRONOUS 'CRYSTALLIZATION' OF DIFFERENT FUNCTIONAL SYSTEMS IN EVOLUTION; IN PARTICULAR: INDEPENDENT INVENTION OF DNA REPLICATION SYSTEMS**

STRUCTURAL GENOMICS:

- 3D STRUCTURE AVAILABLE OR PREDICTABLE FOR THE VAST MAJORITY OF COMMON PROTEIN FAMILIES**

Some important resources for comparative genome analysis

Entrez-Genomes: <http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html>

Clusters of Orthologous Genes: <http://www.ncbi.nlm.nih.gov/COG/>

TIGR database of microbial genome projects:
<http://www.tigr.org/tdb/mdb/hidb/hidb.html>

KEGG: Kyoto Encyclopedia of Genes and Genomes
<http://www.genome.ad.jp/kegg/>

Protein Extraction, Description, and ANalysis Tool at MIPS:
<http://pedant.mips.biochem.mpg.de/>

What Is There (WIT): an environment for interpreting sequenced genomes
<http://wit.mcs.anl.gov/WIT2/>

Acknowledgements

National Center for Biotechnology Information:

- Yuri Wolf, Roman Tatusov, Roland Walker, Michael Galperin, Detlef Leipe, David Lipman

- Department of Pathology, USUHS: Kira Makarova

- Laboratory of Parasitic Diseases, NIAID, NIH:

G. Subramanyan

- UC-Berkeley, UCSF, Stanford: Richard Stephens et al.

(*Chlamydia trachomatis* genome project)